# Méthodes de réduction de dimension pour la construction d'indicateurs de qualité de vie

Amaury Labenne

université de BORDEAUX

Thèse présentée
pour obtenir le grade de

Docteur de
l'université de Bordeaux

École doctorale de Mathématiques et
Informatique

Spécialité : Mathématiques Appliquées

Par Amaury Labenne

# Méthodes de réduction de dimension pour la construction d'indicateurs de qualité de vie

Sous la direction de : Jérôme Saracco
(Co-directrice : Marie Chavent)

Soutenue le 20 novembre 2015 à l'Institut de Mathématiques de Bordeaux

Membres du jury :

| | | | |
|---|---|---|---|
| M. Jérôme Saracco | PROF. | *Bordeaux INP* | Directeur de thèse |
| Mme Marie Chavent | MCF-HDR. | *Université de Bordeaux* | Co-directrice de thèse |
| Mme Nathalie Villa-Vialaneix | MCF-HDR. | *INRA Toulouse* | Rapporteur |
| M. Julien Jacques | PROF. | *Université de Lyon* | Rapporteur |
| Mme Vanessa Kuentz-Simonet | IR. | *IRSTEA Bordeaux* | Examinatrice |
| Mme Tina Rambonilaza | DR. | *IRSTEA Bordeaux* | Examinatrice |
| M. Jean-Michel Poggi | PROF. | *Université Paris Descartes*<br>*LMO, Université Paris-Sud Orsay* | Président du jury |

**Titre :** Méthodes de réduction de dimension pour la construction d'indicateurs de qualité de vie

**Résumé :** L'objectif de cette thèse est de développer et de proposer de nouvelles méthodes de réduction de dimension pour la construction d'indicateurs composites de qualité de vie à l'échelle communale. La méthodologie statistique développée met l'accent sur la prise en compte de la multidimensionnalité du concept de qualité de vie, avec une attention particulière sur le traitement de la mixité des données (variables quantitatives et qualitatives) et l'introduction des conditions environnementales. Nous optons pour une approche par classification de variables et pour une méthode multi-tableaux (analyse factorielle multiple pour données mixtes). Ces deux méthodes permettent de construire des indicateurs composites que nous proposons comme mesure des conditions de vie à l'échelle communale. Afin de faciliter l'interprétation des indicateurs composites construits, une méthode de sélection de variables de type bootstrap est introduite en analyse factorielle multiple. Enfin nous proposons la méthode hclustgeo de classification d'observations qui intègre des contraintes de proximité géographique afin de mieux appréhender la spatialité des phénomènes mis en jeu.

**Mots-clés :** *réduction de dimension, classification de variables, analyses factorielles, méthodes multi-tableaux, données mixtes, indicateurs composites, qualité de vie*

---

**Title :** Dimension reduction methods to construct quality of life indicators

**Abstract :** The purpose of this thesis is to develop and suggest new dimension reduction methods to construct composite indicators on a municipal scale. The developed statistical methodology highlights the consideration of the multi-dimensionality of the quality of life concept, with a particular attention on the treatment of mixed data (quantitative and qualitative variables) and the introduction of environmental conditions. We opt for a variable clustering approach and for a multi-table method (multiple factorial analysis for mixed data). These two methods allow to build composite indicators that we propose as a measure of living conditions at the municipal scale. In order to facilitate the interpretation of the created composite indicators, we introduce a method of selections of variables based on a bootstrap approach. Finally, we suggest the clustering of observations method, named hclustgeo, which integrates geographical proximity constraints in the clustering procedure, in order to apprehend the spatiality specificities better.

**Keywords :** *dimension reduction, variable clustering, factor analysis, multi-table method, mixed data, composite indicators, quality of life*

---

Irstea – Unité de recherche ETBX – Environnement, territoires et infrastructures
50 avenue de Verdun – Gazinet 33612 Cestas Cedex

*"Soyons désinvoltes, n'ayons l'air de rien"*

Noir Désir - Tostaky

# Remerciements

J'ai eu la chance de réaliser ma thèse au sein d'Irstea à Bordeaux qui a financé 50% de ma thèse. Cette thèse s'inscrit également au sein du projet ANR interdisciplinaire ADAPT'EAU [1] qui a financé l'autre moitié de ma thèse. A ce titre, je remercie particulièrement Denis Salles, Directeur de Recherche en sociologie et coordinateur du projet. Je remercie également Frédéric Saudubray, directeur de l'unité ETBX et directeur régional d'Irstea Bordeaux pour m'avoir accueilli et pour avoir mis à ma disposition tous les moyens pour mener à bien ce projet. Au cours de cette thèse, j'ai également passé une partie de mon temps de travail au sein de l'équipe CQFD d'Inria Bordeaux, à ce titre, je remercie Francois Dufour pour m'avoir accueilli au sein de l'équipe. Ce "double" environnement de travail fut très enrichissant, il m'a permis d'une part d'échanger et de collaborer avec des chercheurs de différentes disciplines (économie, sociologie, géographie, écologie, etc.) et d'autre part de développer des méthodes statistiques en lien avec une problématique concrète et actuelle : le changement global.

Je tiens à remercier chaleureusement mes directeurs de thèse Jérôme Saracco et Marie Chavent mais également mon encadrante au sein d'Irstea Vanessa Kuentz-Simonet. Je les remercie tous les trois pour la qualité de leur encadrement, leur disponibilité et leur sympathie mais aussi pour l'intérêt et l'enthousiasme qu'ils ont accordé à mon travail. Leurs encouragements et leur bonne humeur m'ont permis de m'épanouir pendant ces trois années de thèse. Je remercie également Tina Rambonilaza, Directrice de Recherche en économie à Irstea qui a grandement participé à l'encadrement de ce travail, notamment dans la définition du cadre théorique mais également dans l'interprétation des résultats obtenus. Que tous les quatres soient assurés de ma profonde gratitude.

Je tiens également à exprimer toute ma reconnaissance à Nathalie Villa-Vialaneix, Chargée de Recherche à l'Inra de Toulouse et Julien Jacques, Professeur à l'Université

1. ANR ADAPT'EAU. Adaptation aux Variations des Régimes Hydrologiques dans l'Environnement Fluvio-Estuarien de la Garonne-Gironde. Potentialités, mise à l'épreuve et gouvernance d'Options d'Adaptation. Programme Changements Environnementaux Planétaires et Sociétés de l'ANR en 2011 (ANR-11-CEPL-008)

de Lyon, pour avoir accepté d'être les rapporteurs de mon travail de thèse. Je les remercie pour le temps qu'ils ont consacré à la relecture de ce manuscrit.

Merci également à Jean-Michel Poggi, Professeur à l'Université Paris Descartes pour avoir accepté de participer à mon jury de thèse.

Je tiens à remercier l'ensemble de mes collègues de l'unité ETBX d'Irstea. Je ne pourrais pas tous les citer mais certains ont eu un impact direct sur ce travail de thèse. Je pense tout d'abord à Kévin et Baptiste avec qui j'ai collaboré à plusieurs reprises, notamment pour les représentations cartographiques et l'extraction de données, ils ont toujours répondu présent lors de mes multiples sollicitations. Je remercie également Sandrine pour tous les échanges que nous avons eu sur des questions statistiques mais également pour son aide avec LaTeX et R. Ses remarques et suggestions pertinentes ont permis l'amélioration significative des différents packages R développés. Merci aussi à Vincent, mon collègue de bureau, qui nonobstant ses remarques pernicieuses sur l'OM, m'a toujours soutenu et encouragé pendant ces trois années. Je remercie également Stéphanie, Gaby et Maryse qui m'ont accompagné pour les différentes démarches administratives et ont facilité mes déplacements lors des différentes missions effectuées. Pour finir, je remercie Gilou pour son aide relative aux questions informatiques mais également pour les diverses discussions échangées allant du Coty, à la théorie du complot, en passant par hadopi.

Je remercie également mes collègues d'Inria, je pense en particulier à Adrien et Isabelle avec qui j'ai toujours partagé de bons moments que ce soit à Inria ou lors des diverses conférences auxquelles nous avons participé. Merci également à Robin, mon collègue de baby-foot chaud patate.

Merci également aux différentes personnes de l'UFR Sciences et Modélisation de la Victoire. Je pense en premier à Ingrid qui m'a elle aussi aidé pour diverses questions administratives. Merci également à Vincent Couallier et à Brigitte Patouille qui m'ont accompagné lors de mes activités d'enseignement à l'Université.

Je remercie également l'ensemble de mes ami(e)s de Bordeaux et d'ailleurs. Leur présence et leur soutien ont été précieux au cours de ces trois années. J'ai également une pensée émue pour Matthieu qui sans ce fichu caillou aurait été là pour assister à ma soutenance, One Love.

Enfin, mes plus grands remerciements vont à ma famille et à mes parents qui ont fait de moi ce que je suis aujourd'hui. Merci pour leur amour, leur soutien et pour avoir toujours cru en moi quels que soient mes choix. Merci Maman pour tes relectures de ce manuscrit. Pour finir, un grand merci à Drey pour être à mes côtés tous les jours et pour m'avoir soutenu et supporté même dans les moments de doute. Merci de m'avoir laissé être celui-ci ...

# Table des matières

# 1 Présentation générale

## Sommaire



## 1.1 Contexte et enjeux

L'objectif de cette thèse est de développer et de proposer de nouvelles méthodes statistiques pour la construction d'indicateurs composites à l'échelle communale. Ces indicateurs composites seront utilisés pour réaliser des diagnostics à l'échelle des territoires afin d'analyser leur vulnérabilité dans un contexte de changements globaux (changement climatique, évolution de la localisation des activités économiques, modifications de la structure des populations, etc.).

Le concept de développement durable est mobilisé dans le sens où il désigne un territoire sur lequel s'articule des phénomènes socio-économiques et environnementaux. Dans ce cadre, nous rejoignons les travaux en cours sur la construction d'indicateurs de développement durable. Ces indicateurs de développement durable sont censés répondre à trois grandes fonctions : scientifique d'abord afin d'objectiver les progrès en matière de développement durable ; politique ensuite car il est nécessaire d'identifier les priorités en matière d'actions publiques et d'évaluer leurs performances ; sociétale enfin pour faciliter la communication auprès du public. Afin de donner une plus grande consistance à ces indicateurs, on assiste, depuis le début des années 90, à une profusion d'initiatives,

émanant de chercheurs, d'associations, d'institutions statistiques, ou d'organisations internationales pour la construction d'indicateurs de développement durable. En France, récemment, Bovar and Nirascou (2010) élaborent une liste d'une quarantaine d'indicateurs simples de développement durable territoriaux. Leur démarche consiste à mettre à disposition pour chaque pilier du développement durable (économique, social, environnemental) un ensemble d'indicateurs d'états ou de progrès qui servent de référence et d'exemples concrets pour les acteurs locaux et qui sont mis à jour régulièrement. Cependant, il est difficile d'identifier des tendances parmi un ensemble d'indicateurs simples, d'où la nécessité de construire des indicateurs composites. On appelle indicateur composite un indicateur rassemblant plusieurs indicateurs simples (ou variables). Le but d'un indicateur composite est de pouvoir mesurer des concepts multidimensionnels qui ne peuvent pas être capturés par un indicateur simple. Le développement de tels indicateurs s'explique par le besoin de disposer d'une information simple, facile à retenir ou à communiquer et qui permet de faire des comparaisons ou d'établir des palmarès entre pays ou régions.

Dans l'optique de construire des indicateurs composites à des fins de diagnostic des territoires, nous utilisons la notion de qualité de vie. En effet, le cadre analytique fourni par le Millenium Ecosystem Assessment (MEA), prolongé récemment par l'initiative CICES, voir Haines-Young and Potschin (2010), permet de questionner les liens entre les différentes dimensions de la qualité de vie (conditions de logement, conditions d'emplois, environnement naturel, etc.) et les facteurs qui impactent son évolution. Ainsi, si l'objectif est de décrire la situation socio-économique et environnementale des territoires et leur évolution, la mesure de la qualité de vie et de ses différentes dimensions constitue un indicateur pertinent pour l'évaluation des états des sociétés. Cette notion permet d'intégrer le double aspect socio-économique et environnemental dans l'évaluation, tout en explicitant les différentes dimensions explicatives de cette qualité de vie : les conditions de vie comme les conditions environnementales, politiques et sociales, voir Costanza et al. (2007). La combinaison des variables associées aux différentes dimensions de la qualité de vie fait alors écho aux méthodes statistiques de réduction de dimension qui semblent bien adaptées pour la construction d'indicateurs composites. L'objectif de ces méthodes est de résumer l'information apportée par un ensemble de variables (ou indicateurs simples) à l'aide d'un ensemble réduit de nouvelles variables qui seront le plus possible liées aux variables initiales. Dans la littérature, la méthode la plus souvent utilisée pour construire des indicateurs composites de qualité de vie est l'analyse en composantes principales (ACP), voir par exemple, Wood et al. (2010), Haq and Zia (2013), Krishnan (2014) ou Reynard and Vialette (2014).

Dans cette thèse, nous proposons de dépasser la simple utilisation de l'ACP et de mobiliser des méthodes de réduction de dimension innovantes afin d'appréhender la

multidimensionnalité du concept de qualité de vie.

## 1.2 Les indicateurs composites proposés dans cette thèse

La construction d'indicateurs composites soulève plusieurs questions d'ordre méthodologique. Quelles variables inclure dans la construction de l'indicateur ? Comment les agréger et quelles pondérations leur attribuer ? L'Organisation de Coopération et de Développement Économiques [1] (OCDE) et le "Joint Research Center" [2] (JRC) ont publié un ouvrage, voir Nardo (2008), dont le but est de fournir un cadre et des règles nécessaires à la construction d'indicateurs composites. Bien que ce guide se concentre sur les indicateurs composites à l'échelle des pays, l'ensemble des bonnes pratiques expliquées peuvent être reproduites pour construire les indicateurs composites à l'échelle communale.

### 1.2.1 Deux règles importantes pour la construction d'indicateurs composites

Dans cette thèse, nous mettons l'accent sur deux points majeurs que soulève la construction d'indicateurs composites : la définition d'un cadre théorique rigoureux et le développement d'une méthodologie pour l'agrégation et la pondération des variables.

**Le cadre théorique et la sélection des variables.** La première règle dans la construction d'indicateurs composites est la définition d'un cadre théorique. Ce cadre théorique, souvent défini à l'aide d'experts du domaine, sert de base à la sélection des variables utilisées pour construire les indicateurs composites en conformité avec l'objectif souhaité. Le cadre théorique de notre étude est basé sur le concept de mesure de la qualité de vie en lien avec la vulnérabilité des territoires. Pour définir un cadre théorique de la mesure de la qualité de vie, nous prenons comme référence l'architecture du "Système Européen d'Indicateurs Sociaux", qui couvre initialement 13 dimensions de la vie, voir Noll (2002). Cependant, étant donné la problématique traitée, liée à la vulnérabilité des territoires, certaines dimensions peuvent être considérées comme moins importantes (par exemple : les loisirs, la culture, la religion et la spiritualité) tandis que d'autres devraient avoir plus d'importance. C'est le cas des opportunités du marché du travail, des conditions de travail, ou encore de l'accès aux différents services

1. OCDE. `http://www.oecd.org/fr/`
2. JRC. `http://ec.europa.eu/jrc/`

de santé, qui peuvent jouer un rôle capital pour la capacité d'adaptation des communautés. C'est également le cas pour l'environnement social des individus, l'interaction sociale et les modes de vie. Enfin, il y a d'autres dimensions de la vie telles que les conditions environnementales et l'accès aux aménités naturelles qui méritent une attention particulière. Ainsi, les dimensions suggérées pour être couvertes par les indicateurs de qualité de vie que nous allons construire sont les suivantes : (1) les conditions de logement, (2) l'état du marché du travail et l'accès à l'emploi, (3) les conditions financières des ménages, (4) l'accès à l'éducation, (5) l'accès aux services de santé, (6) l'accessibilité et la qualité des autres services, (7) l'état de l'environnement social et (8) l'état de l'environnement naturel. Ce cadre théorique et les dimensions associées nous serviront de base pour la sélection des variables à inclure dans les indicateurs composites que nous allons proposer.
Une fois le cadre théorique défini, il est nécessaire de sélectionner les indicateurs simples (variables) à inclure dans la construction de l'indicateur composite selon leur pertinence pour le phénomène mesuré, leur mesurabilité, mais aussi selon leur disponibilité. Cela pose la question suivante : pour chaque dimension de la qualité de vie présentée précédemment, comment choisir la ou les variables les plus pertinentes pour la construction d'indicateurs ? Afin de limiter la subjectivité et dans l'objectif de capter les différents aspects de la qualité de vie, nous choisissons d'utiliser l'ensemble des variables à notre disposition au sein de chaque dimension.

**Agrégation et pondération des variables.** L'approche consistant à utiliser l'ensemble des variables à notre disposition amène la problématique de l'agrégation et de l'attribution d'un poids à chacune des variables utilisées. La question qui se pose alors est de savoir comment appréhender la combinaison des différentes variables pour construire des indicateurs composites tout en évitant le problème de redondance de l'information. L'enjeu de la thèse se situe à ce niveau : comment agréger et pondérer les différentes variables pour rendre compte de la multidimensionnalité du concept de la qualité de vie. A ce titre, nous proposons dans ce travail de thèse des méthodes de réduction de dimension innovantes permettant d'apporter une réponse à cette question. L'objectif de ces méthodes est de résumer un ensemble initial de variables par un petit nombre de nouvelles variables, tout en reconstruisant le maximum d'information contenue dans les données originales, elles permettent également de surmonter la question du choix des variables à sélectionner en intégrant différentes variables éventuellement fortement liées entre elles. D'une part ces méthodes permettent de traiter la multidimensionnalité des phénomènes mesurés et d'autre part, elles permettent de considérer les étapes d'agrégation et de pondération des variables de manière simultanée. Ainsi les coefficients associés aux différentes variables dans les indicateurs composites ne se-

ront pas définis a priori ou à dire d'experts mais ils seront issus des relations entre les variables.

### 1.2.2 Les données utilisées

Afin de construire des indicateurs composites de qualité de vie, il est important d'utiliser des données pertinentes qui reflètent les variations de cette qualité de vie sur une période donnée. Nous avons vu précédemment les dimensions de la qualité de vie que nous allons utiliser pour sélectionner les variables à inclure dans la construction des indicateurs composites. De plus, l'information disponible doit être aussi comparable que possible sur une période donnée. Dans cette optique, nous avons utilisé deux jeux de données relatifs à deux dates différentes (1999 et 2009). Les données socio-économiques ont été extraites via le portail SIDDT de l'IRSTEA de Grenoble [3] qui rassemble les données de recensement publiées par l'Institut National de Statistique et des Études Économiques [4] (INSEE). Les données relatives à l'usage des sols proviennent de la base Corine Land Cover (CLC) avec une précision de 25 hectares. La base CLC de 2000 est utilisée pour l'année 1999 et la base CLC de 2006 est utilisée pour l'année 2009. La combinaison de ces différentes bases de données fournit un système de 55 variables pour l'année 1999. Pour l'année 2009, les variables correspondantes sont au nombre de 46.
De plus les données utilisées sont "mixtes", c'est à dire que l'on dispose de variables quantitatives (ou numériques) mais également de variables qualitatives (ou catégorielles). La mixité des données implique d'utiliser des méthodes statistiques adaptées prenant en compte ces deux types de variables.

### 1.2.3 Les méthodes statistiques proposées

**La classification de variables.** La première approche retenue pour la construction d'indicateurs composites présentée dans cette thèse est l'approche par classification de variables (appelée également clustering de variables) incluse dans le package R `ClustOfVar`. Cette méthode développée par Chavent et al. (2012) semble être pertinente pour la construction d'indicateurs composites. En effet, elle permet d'agréger des variables en clusters homogènes sans aucun a priori sur la structuration des données, c'est à dire que la structuration en groupes de variables (appartenant aux différentes dimensions de la qualité de vie) n'est pas prise en compte et toutes les variables participent de la même façon à la construction des indicateurs. Le but de la méthode est de construire des clusters contenant des variables fortement liées entre elles. Cela

3. Système d'Information Dédié aux Territoires. SIDDT. `http://siddt.irstea.fr/`
4. Institut national de la statistique et des études économiques. INSEE. `http://www.insee.fr/`

nous permettra de comprendre quelles variables se ressemblent et apportent la même information. De plus, pour chaque cluster de variables, la méthode construit une nouvelle variable (appelée variable synthétique) qui est la plus liée possible aux variables du cluster. Cette variable synthétique est une combinaison linéaire des variables du cluster auquel elle est associée. Les différentes variables synthétiques ainsi construites seront de potentiels indicateurs composites de qualité de vie. Le package `ClustOfVar` a d'abord été utilisé dans sa version originale, puis nous l'avons amélioré en ajoutant certaines sorties numériques et graphiques afin de faciliter l'interprétation des indicateurs obtenus.

**L'analyse factorielle multiple appliquée aux données mixtes.** La seconde approche utilisée pour la construction d'indicateurs composites est l'analyse factorielle multiple (AFM). Cette méthode d'analyse factorielle est spécialement dédiée au cas où des observations sont décrites par plusieurs groupes de variables (les différentes dimensions de la qualité de vie vues précédemment). Elle peut être vue comme une ACP réalisée sur plusieurs tableaux de données. Les méthodes dites multi-tableaux comme l'AFM permettent d'équilibrer l'information apportée par les différents groupes de variables afin que la construction des composantes principales ne soit pas trop influencée par des groupes ayant une forte structure ou un grand nombre de variables. Cependant, l'écriture actuelle de l'AFM ne permettait pas de prendre en compte les groupes contenant des variables quantitatives et des variables qualitatives. Du fait de la mixité de nos données, nous avons développé une extension de l'AFM, appelée MFAmix, qui prend en compte les variables quantitatives et les variables qualitatives. Cette méthode semble adaptée pour traiter le problème de multidimensionnalité apporté par les différentes dimensions de la qualité de vie. Les composantes principales obtenues peuvent être vues comme de nouvelles variables qui résument "au mieux" les variables incluses dans l'analyse. De plus, ces composantes principales peuvent s'écrire comme une combinaison linéaire pondérée des variables d'origine. Ce sont donc de potentiels indicateurs composites de qualité de vie.

**La classification d'observations.** Une fois les indicateurs composites de qualité de vie construits (quelle que soit la méthode utilisée), il est important de regarder les valeurs de ces indicateurs sur les différentes communes étudiées afin d'effectuer un diagnostic des territoires. Plutôt que de comparer les communes une à une, nous avons choisi de les rassembler en classes homogènes afin d'effectuer des comparaisons entre classes de communes mais aussi afin de repérer facilement les communes ayant les mêmes profils. Les classes créées doivent être de telle sorte qu'au sein d'une même classe les communes se ressemblent le plus possible mais également que les communes soient

le plus différentes possible si elles sont dans des classes différentes. Plusieurs méthodes de classification d'observations existent : classification ascendante hiérarchique (CAH), méthode de type k-means, etc. Cependant, nous privilégions la CAH car elle ne nécessite pas de définir un nombre de classes au préalable. Ainsi, une fois que nous aurons construits des indicateurs sur une zone d'étude définie, nous réaliserons des typologies des communes à partir de leurs valeurs (scores) sur les indicateurs composites.

**La spatialité des phénomènes.** Une fois les typologies de communes réalisées, il est intéressant de réaliser une cartographie des classes obtenues, ceci étant une première étape pour explorer la spatialité des phénomènes. Cependant, dans notre cas, la classification est effectuée sur des observations spatialisées (les communes) et lorsque l'on utilise des méthodes de classification classiques, comme la CAH avec critère de Ward, il arrive que la typologie obtenue soit assez fragmentée géographiquement lorsqu'on la représente sur une carte. Cependant, il semble judicieux de penser que deux communes contigües subissent a priori des influences semblables mais non mesurées (pression d'urbanisation, développement local, etc) et auront donc tendance à se ressembler. En effet, il peut être nécessaire de diminuer l'importance de certaines variations locales qui peuvent avoir un caractère partiellement aléatoire et dégager ainsi des tendances générales sur lesquelles les zones proches géographiquement auront tendance à se ressembler. Il est donc important d'introduire de l'information spatiale (ou géographique) sur les communes étudiées. Nous avons choisi d'intégrer cette information dans la procédure de classification des communes réalisée sur les indicateurs composites. La méthode hclustgeo développée dans ce but permet de réaliser une CAH prenant en compte l'information apportée par les indicateurs composites mais également les distance géographiques entre les communes.

## 1.3 Plan de la thèse

Pour commencer, nous allons présenter au Chapitre 2 la méthode d'analyse factorielle de données mixtes appelée PCAmix. Nous verrons par la suite que cette méthode est au coeur de ce travail de thèse. Ensuite, le Chapitre 3 sera dédié à la présentation et à l'utilisation d'une méthode de classification de variables pour la construction d'indicateurs composites. Par la suite, la méthode MFAmix ainsi qu'une application de celle-ci sera présentée au Chapitre 4. Pour finir, le Chapitre 5 présente la méthode hclustgeo qui prend en compte les distances géographiques entre les communes dans la procédure de classification. Chaque chapitre est organisé de la manière suivante : tout d'abord une présentation théorique de la méthode est donnée, ensuite la méthode est illustrée sur un exemple simple à l'aide de packages R spécifiques, des extraits de code

seront inclus afin de comprendre comment utiliser la méthode et comment interpréter les résultats obtenus. Pour finir une section sera dédiée à l'utilisation de la méthode sur un exemple concret dans le but de construire des indicateurs composites et d'effectuer des typologies des communes étudiées.

# 2

# Analyse factorielle de données mixtes : la méthode PCAmix

## Sommaire



## 2.1 Introduction

Nous présentons dans ce chapitre la méthode d'analyse factorielle mixte appelée PCAmix dans la suite. Cette méthode peut être vue comme un mélange de l' analyse en composantes principales (ACP) (pour des variables quantitatives) et de l'analyse des correspondances multiples (ACM) (pour des variables qualitatives). Différentes méthodes d'analyse factorielle de données mixtes ont été proposées, voir par exemple Escofier (1979), Kiers (1991), Pagès (2004) ou Saporta (2006). La méthode PCAmix que nous proposons ici est basée sur la méthode proposée par Chavent et al. (2012). Nous allons réécrire cette méthode à l'aide d'une décomposition en valeurs singulières généralisée (GSVD). La méthode PCAmix joue un rôle prépondérant dans ce travail

de thèse car elle est au cœur des méthodes de classification de variables et d'analyse factorielle multiple qui seront présentées dans les chapitres suivants. La Section 2.2 présente le principe de la GSVD et montre comment elle peut être utilisée pour réécrire l'ACP et l'ACM. La Section 2.3 explique la mise en œuvre de la méthode PCAmix et les différentes sorties nécessaires à l'interprétation des résultats. Finalement, la Section 2.4 décrit l'utilisation de PCAmix à l'aide du package `PCAmixdata` sur un exemple simple, des extraits de code R relatifs aux exemples sont inclus.

## 2.2 GSVD et Analyses factorielles

Nous allons voir dans cette section comment la décomposition en valeurs singulières généralisée peut être utilisée pour l'écriture de différentes méthodes d'analyses factorielles. Après avoir rappelé le principe de la GSVD, nous montrerons comment l'ACP standardisée et l'ACM peuvent être écrites à l'aide de différentes GSVD.

### 2.2.1 GSVD d'une matrice numérique Z

La GSVD permet une décomposition d'une matrice $\mathbf{Z}$, de dimension $n \times p$, en utilisant deux matrices carrées définies positives $\mathbf{N}$ et $\mathbf{M}$, où $\mathbf{N}$ est une métrique sur $\mathbb{R}^n$ et $\mathbf{M}$ est une métrique sur $\mathbb{R}^p$. La GSVD de $\mathbf{Z}$ avec les métriques $\mathbf{N}$ et $\mathbf{M}$ donne la décomposition suivante :

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t, \tag{2.2.1}$$

où

- $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ est la matrice diagonale, de dimension $r \times r$, des valeurs singulières de $\mathbf{ZMZ}^t\mathbf{N}$ et $\mathbf{Z}^t\mathbf{NZM}$, où $r$ est le rang de $\mathbf{Z}$,
- $\mathbf{U}$ est la matrice $n \times r$ des $r$ premiers vecteurs propres de $\mathbf{ZMZ}^t\mathbf{N}$ avec $\mathbf{U}^t\mathbf{NU} = \mathbb{I}_r$ où $\mathbb{I}_r$ est la matrice identité de dimension $r$,
- $\mathbf{V}$ est la matrice $p \times r$ des $r$ premiers vecteurs propres de $\mathbf{Z}^t\mathbf{NZM}$ avec $\mathbf{V}^t\mathbf{MV} = \mathbb{I}_r$.

La GSVD de $\mathbf{Z}$ est obtenue en réalisant une décomposition en valeurs singulières classique (SVD) de la matrice $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{ZM}^{1/2}$, ce qui est équivalent à une GSVD avec les métriques $\mathbb{I}_n$ sur $\mathbb{R}^n$ et $\mathbb{I}_p$ sur $\mathbb{R}^p$. Cela donne :

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^t. \tag{2.2.2}$$

On procède de la manière suivante pour obtenir les matrices $\mathbf{U}$, $\mathbf{\Lambda}$ et $\mathbf{V}$ :

$$\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}, \quad \mathbf{U} = \mathbf{N}^{-1/2}\tilde{\mathbf{U}}, \quad \mathbf{V} = \mathbf{M}^{-1/2}\tilde{\mathbf{V}}. \tag{2.2.3}$$

### 2.2.2 ACP de Z avec métriques

La GSVD peut être utilisée pour introduire des poids sur les lignes et les colonnes de $\mathbf{Z}$. Les métriques $\mathbf{N}$ et $\mathbf{M}$ associées à la GSVD sont alors les matrices diagonales de ces poids. Les coordonnées factorielles des lignes (i.e des observations) et les coordonnées factorielles des colonnes (i.e des variables) sont obtenues de la manière suivante.

**Coordonnées factorielles des lignes.** On note $\mathbf{F}$ la matrice $n \times r$ des coordonnées factorielles des lignes de $\mathbf{Z}$. Ces coordonnées correspondent à la projection orthogonale, associée à la matrice de produit scalaire $\mathbf{M}$, des $n$ lignes de $\mathbf{Z}$ sur les composantes principales (aussi appelées axes factoriels) engendrées par les vecteurs $\mathbf{v}_1, \ldots, \mathbf{v}_r$ (vecteurs colonnes de $\mathbf{V}$). On a donc :

$$\mathbf{F} = \mathbf{ZMV}. \tag{2.2.4}$$

On déduit de (2.2.1) que :

$$\mathbf{F} = \mathbf{ZMV} = \mathbf{U\Lambda V}^t\mathbf{MV} = \mathbf{U}\Lambda. \tag{2.2.5}$$

On rappelle que les colonnes de $\mathbf{V}$ sont les vecteurs propres de $\mathbf{Z}^t\mathbf{NZM}$ qui peuvent être obtenues en résolvant la séquence suivante (indexée par $k$) de problèmes d'optimisation :

$$\begin{array}{lll} \text{maximiser} & \|\mathbf{ZMv}_k\|^2_{\mathbf{N}} & \\ \text{tel que} & \mathbf{v}_k^t\mathbf{Mv}_l = 0 \quad \forall 1 \leq l < k, & \\ & \mathbf{v}_k^t\mathbf{Mv}_k = 1, & \end{array} \tag{2.2.6}$$

avec $\|\mathbf{x}\|^2_{\mathbf{N}} = \mathbf{x}^t\mathbf{Nx}$. On note $\mathbf{f}_k = \mathbf{ZMv}_k$ une colonne de $\mathbf{F}$. Les vecteurs $\mathbf{v}_1, \ldots, \mathbf{v}_r$ sont définis de manière à ce que $\|\mathbf{f}_k\|^2_{\mathbf{N}} = \lambda_k$ soit maximal. Les colonnes de $\mathbf{F}$ sont appelées composantes principales.

**Coordonnées factorielles des colonnes.** On note $\mathbf{A}$ la matrice $p \times r$ des coordonnées factorielles des colonnes. Ces coordonnées factorielles correspondent à la projection orthogonale, associée à la matrice de produit scalaire $\mathbf{N}$ des $p$ colonnes sur les axes engendrés par les vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_r$ (colonnes de $\mathbf{U}$). Cette définition donne :

$$\mathbf{A} = \mathbf{Z}^t\mathbf{NU}. \tag{2.2.7}$$

On déduit de (2.2.1) que :

$$\mathbf{A} = \mathbf{V}\Lambda. \tag{2.2.8}$$

On rappelle que les colonnes de $\mathbf{U}$ sont les vecteurs propres de $\mathbf{ZMZ}^t\mathbf{N}$ et peuvent être obtenues en résolvant la séquence suivante (indexée par $k$) de problèmes d'optimisation :

$$\begin{array}{lll} \text{maximiser} & \|\mathbf{Z}^t\mathbf{Nu}_k\|^2_{\mathbf{M}} & \\ \text{tel que} & \mathbf{u}_l^t\mathbf{Nu}_k = 0 \quad \forall 1 \leq l < k, & \\ & \mathbf{u}_k^t\mathbf{Nu}_k = 1. & \end{array} \tag{2.2.9}$$

On note $\mathbf{a}_k = \mathbf{Z}^t\mathbf{N}\mathbf{u}_k$ une colonne de $\mathbf{A}$. Les vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_r$ sont donc définis de telle sorte que $\|\mathbf{a}_k\|^2_{\mathbf{M}} = \lambda_k$ est maximal. Les colonnes de $\mathbf{A}$ sont appelées les "loadings".

Notez que $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}$ dans la SVD standard de $\tilde{\mathbf{Z}} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{M}^{1/2}$ dans (2.2.2). Cela donne :

$$\lambda_k = \|\mathbf{a}_k\|^2_{\mathbf{M}} = \|\tilde{\mathbf{a}}_k\|^2_{\mathbb{I}_p},$$

où $\tilde{\mathbf{a}}_k$ est la $k$-ème colonne de $\widetilde{\mathbf{A}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}$.

### 2.2.3 ACP et ACM standardisées

Cette section présente comment l'ACP classique (pour données quantitatives) et l'ACM classique (pour données qualitatives) peuvent être obtenues à partir de la GSVD des matrices $\mathbf{Z}$, $\mathbf{N}$ et $\mathbf{M}$. La matrice numérique $\mathbf{Z}$ est obtenue en recodant d'une certaine manière la matrice $\mathbf{X}$ de données d'origine, la métrique $\mathbf{N}$ (resp. $\mathbf{M}$) étant la matrice diagonale des poids des lignes (resp. des colonnes) de $\mathbf{Z}$.

**ACP standardisée.** La matrice de données $\mathbf{X}$ de dimension $n \times p$ sur laquelle on veut réaliser l'ACP contient $n$ observations décrites par $p$ variables quantitatives. L'étape de recodage consiste à centrer et réduire les colonnes de $\mathbf{X}$ pour construire la matrice standardisée $\mathbf{Z}$ (ainsi $\frac{1}{n}\mathbf{Z}^t\mathbf{Z}$ correspond à la matrice des corrélations). Les $n$ lignes (observations) sont pondérées par $\frac{1}{n}$ et les $p$ colonnes (variables quantitatives) sont pondérées par 1 (toutes les variables ont le même poids). Cela donne $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$ et $\mathbf{M} = \mathbb{I}_p$. La métrique $\mathbf{M}$ indique que la distance entre deux observations est la distance euclidienne classique entre deux lignes de $\mathbf{Z}$. Ainsi, l'inertie totale de $\mathbf{Z}$ est égale à $p$. La matrice $\mathbf{F}$ des coordonnées factorielles des lignes et la matrice $\mathbf{A}$ des loadings (coordonnées des variables) sont directement calculées à partir de (2.2.5) et (2.2.8). Les propriétés classiques de l'ACP sont les suivantes :

- Chaque coordonnée factorielle $a_{jk}$ correspond à la corrélation linéaire entre la variable quantitative $\mathbf{x}_j$ (la $j$-ème colonne de $\mathbf{X}$)et la $k$-ème composante principale $\mathbf{f}_k$ (la $k$-ème colonne de $\mathbf{F}$) :

$$a_{jk} = \mathbf{z}_j^t\mathbf{N}\mathbf{u}_k = r(\mathbf{x}_j, \mathbf{f}_k), \tag{2.2.10}$$

  avec $\mathbf{u}_k = \frac{\mathbf{f}_k}{\lambda_k}$ et $\mathbf{z}_j$ (resp. $\mathbf{x}_j$ ) la $j$-ème colonne de $\mathbf{Z}$ (resp. $\mathbf{X}$).
- Chaque valeur propre (carré de la valeur singulière) $\lambda_k$ correspond à la variance de la $k$-ème composante principale :

$$\lambda_k = \|\mathbf{f}_k\|^2_{\mathbf{N}} = \mathrm{Var}(\mathbf{f}_k). \tag{2.2.11}$$

- De plus, chaque valeur propre $\lambda_k$ est aussi la somme des corrélations au carré entre les $p$ variables et la $k$-ème composante principale :

$$\lambda_k = \|\mathbf{a}_k\|^2_{\mathbf{M}} = \sum_{j=1}^{p} r^2(\mathbf{x}_j, \mathbf{f}_k). \tag{2.2.12}$$

**ACM standardisée.** La matrice de données $\mathbf{X}$ de dimension $n \times p$ sur laquelle on veut réaliser l'ACM contient $n$ observations décrites par $p$ variables qualitatives. Chaque variable $j$ possède $m_j$ modalités (aussi appelées "levels" dans le package `PCAmixdata`), on note $m$ le nombre total de modalités des $j$ variables. L'étape de recodage de la matrice $\mathbf{X}$ consiste à recoder chaque modalité en variable binaire et ainsi construire la matrice d'indicatrices $\mathbf{G}$ de dimension $n \times m$. Habituellement, l'ACM est réalisée en appliquant l'analyse des correspondances sur la matrice $\mathbf{G}$. En analyse des correspondances les coordonnées factorielles des lignes (resp. des colonnes) sont obtenues en réalisant une ACP sur deux matrices différentes : la matrice des profils lignes (resp. la matrice des profils colonnes). Ici, nous présentons une méthode pour calculer les coordonnées factorielles (des lignes et des colonnes) de l'ACM à l'aide d'une seule ACP avec métriques.

On note maintenant $\mathbf{Z}$ la matrice des indicatrices $\mathbf{G}$ centrée. Les $n$ lignes (observations) sont pondérées par $\frac{1}{n}$ et les $m$ colonnes (modalités des variables qualitatives) sont pondérées par $\frac{n}{n_s}$, l'inverse de la fréquence de la modalité $s$, où $n_s$ désigne le nombre d'observations possédant la modalité $s$. Cela donne $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$ et $\mathbf{M} = \text{diag}(\frac{n}{n_s}, s = 1\ldots, m)$. La métrique $\mathbf{M}$ indique que la distance entre deux observations est une distance euclidienne pondérée semblable à la distance du $\chi^2$ utilisée en analyse des correspondances. Cette distance accorde une plus grande importance aux modalités rares. L'inertie totale de $\mathbf{Z}$ avec cette distance et les poids $\frac{1}{n}$ sur les observations est égale à $p - m$. La GSVD de $\mathbf{Z}$ avec ces deux métriques permet de calculer directement la matrice $\mathbf{F}$ des coordonnées factorielles des lignes à partir de (2.2.5). Cependant, les coordonnées factorielles des modalités ne sont pas calculées directement à partir de (2.2.8). En notant $\mathbf{A}^*$ la matrice des coordonnées factorielles des modalités, on a :

$$\mathbf{A}^* = \mathbf{MA} = \mathbf{MV\Lambda}. \tag{2.2.13}$$

On notera que ce résultat est différent du résultat de l'ACP avec métriques où les coordonnées factorielles des colonnes sont donnés par : $\mathbf{A} = \mathbf{V\Lambda}$. La preuve de ce résultat est donnée en Annexe A.

Les propriétés classiques de l'ACM sont les suivantes :

- Chaque coordonnée factorielle $a^*_{sk}$ est la valeur moyenne (normalisée) des coordonnées des observations possédant la modalité $s$.

$$a^*_{sk} = \frac{n}{n_s} a_{sk} = \frac{n}{n_s} \mathbf{z}_s^t \mathbf{N} \mathbf{u}_k = \bar{u}^s_k, \tag{2.2.14}$$

avec $\mathbf{z}_s$ la $s$-ème colonne de $\mathbf{Z}$ et $\bar{u}_k^s$ la valeur moyenne de $\mathbf{u}_k$ calculée sur les observations possédant la modalité $s$.

- Chaque valeur propre $\lambda_k$ est la somme des rapports de corrélation entre les $p$ variables qualitatives et la $k$-ème composante principale (qui est numérique) :

$$\lambda_k = \|\mathbf{a}_k\|_{\mathbf{M}}^2 = \|\mathbf{a}_k^*\|_{\mathbf{M}^{-1}}^2 = \sum_{j=1}^{p} \eta^2(\mathbf{f}_k|x_j). \tag{2.2.15}$$

Le rapport de corrélation $\eta^2(\mathbf{f}_k|x_j)$ mesure la part de variance de $\mathbf{f}_k$ expliquée par la variable qualitative $j$.

Effectuer l'ACM de cette façon implique quelques changements mineurs par rapport à l'ACM classique calculée en effectuant une analyse des correspondances de la matrice des indicatrices $\mathbf{G}$ :

- L'inertie totale est multipliée par $p$ et est donc égale à $m - p$. Cette propriété sera utile pour la méthode PCAmix afin d'équilibrer l'inertie apportée par les variables quantitatives (égale au nombre de variables quantitatives) et l'inertie des variables qualitatives (maintenant égale au nombre de modalités moins le nombre de variables qualitatives).
- Les coordonnées des modalités restent inchangées. Cependant les valeurs propres sont multipliées par $p$ et donc les coordonnées des observations sont multipliées par $\sqrt{p}$.

## 2.3 La méthode PCAmix

On se place ici dans le cas où le tableau de données $\mathbf{X}$ que l'on veut analyser comporte $n$ observations décrites par $p_1$ variables quantitatives et $p_2$ variables qualitatives. La matrice $\mathbf{X}$ est vue comme la concaténation de deux matrices $\mathbf{X}_1$ et $\mathbf{X}_2$. La matrice $\mathbf{X}_1$ de dimension $n \times p_1$ contient la description des $n$ observations par les $p_1$ variables quantitatives et la matrice $\mathbf{X}_2$ de dimension $n \times p_2$ contient la description des mêmes observations par les $p_2$ variables qualitatives. On note $m$ le nombre total de modalités des $p_2$ variables qualitatives.

### 2.3.1 Algorithme de PCAmix

La méthode PCAmix est une procédure en deux étapes qui peut être vue comme un mélange de l'ACP et de l'ACM. La procédure est la suivante :

**Etape 1 : Recodage des matrices $\mathbf{X}_1$ et $\mathbf{X}_2$.**

1. Construire la matrice réelle $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$ de dimension $n \times (p_1 + m)$ où :

↪ $\mathbf{Z}_1$ est la matrice $\mathbf{X}_1$ centrée et standardisée (comme en ACP),
↪ $\mathbf{Z}_2$ est la matrice centrée $\mathbf{G}$ des indicatrices de $\mathbf{X}_2$ (comme en ACM).

2. Construire la matrice diagonale $\mathbf{N}$ des poids des lignes de $\mathbf{Z}$. Les $n$ lignes sont pondérées par $\frac{1}{n}$, ainsi $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$.
3. Construire la matrice diagonale $\mathbf{M}$ des poids des colonnes de $\mathbf{Z}$.
   ↪ Les $p_1$ premières colonnes sont pondérées par 1 (comme en ACP).
   ↪ Les $m$ dernières colonnes sont pondérées par $\frac{n}{n_s}$ (comme en ACM), où $n_s$ est le nombre d'observations possédant la modalité $s$.

La métrique $\mathbf{M}$ indique que la distance entre deux lignes de $\mathbf{Z}$ est un mélange entre la distance euclidienne classique utilisée en ACP (pour les $p_1$ premières colonnes) et la distance semblable à celle du $\chi^2$ utilisée en ACM (pour les $m$ dernières colonnes). Ainsi, l'inertie totale de $\mathbf{Z}$ avec cette distance et les poids $\frac{1}{n}$ sur les observations est égale à $p_1 + m - p_2$.

**Etape 2 : Obtention des coordonnées factorielles des lignes et des colonnes.**

1. La GSVD de $\mathbf{Z}$ avec les métriques $\mathbf{N}$ et $\mathbf{M}$ donne la décomposition suivante :

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t,$$

   comme cela est détaillé à la Section 2.2.

2. Les coordonnées factorielles des lignes ($n$ observations) sont obtenues comme suit :

$$\mathbf{F} = \mathbf{ZMV}, \tag{2.3.1}$$

   ou directement à partir de la GSVD :

$$\mathbf{F} = \mathbf{U}\mathbf{\Lambda}. \tag{2.3.2}$$

3. Les coordonnées factorielles des colonnes ($p_1$ variables quantitatives et $m$ modalités) sont obtenues comme suit :

$$\mathbf{A}^* = \mathbf{MV}\mathbf{\Lambda}. \tag{2.3.3}$$

   La matrice $\mathbf{A}^*$ est découpée comme suit : $\mathbf{A}^* = \begin{bmatrix} \mathbf{A}_1^* \\ \mathbf{A}_2^* \end{bmatrix} \begin{matrix} \} p_1 \\ \} m \end{matrix}$ où
   ↪ $\mathbf{A}_1^*$ contient les coordonnées factorielles des $p_1$ variables quantitatives,
   ↪ $\mathbf{A}_2^*$ contient les coordonnées factorielles des $m$ modalités.

On remarque que le calcul des coordonnées factorielles des colonnes dans PCAmix est légèrement différent du calcul effectué dans l'ACP avec métriques où les coordonnées

des colonnes sont donnés par $\mathbf{A} = \mathbf{V\Lambda}$. Cela est dû au fait que l'ACM n'est pas une simple ACP comme expliqué en Section 2.2.3.

Il est important de noter que, pour des données quantitatives, PCAmix est exactement une ACP standardisée et pour des données qualitatives, PCAmix est équivalent à l'ACM.

### 2.3.2 Sorties numériques de PCAmix

La méthode PCAmix permet de représenter le lien entre les variables, les observations et les composantes principales (ou axes factoriels) en les représentant sur divers graphiques grâce à leurs coordonnées factorielles. De plus, les propriétés utilisées en ACP et en ACM pour interpréter les plans factoriels restent vraies avec PCAmix.

— La coordonnée factorielle $a^\star_{jk}$ d'une variable quantitative $j$ sur un axe $k$ est égale à la corrélation entre la variable $j$ et la composante principale $k$ ($k$-ème colonne de $\mathbf{F}$).

— Comme en ACM, la coordonnée factorielle $a^\star_{sk}$ d'une modalité $s$ sur un axe $k$ correspond à la valeur moyenne des coordonnées factorielles (standardisée) des observations ($f_{ik}$) possédant la modalité $s$.

En plus des coordonnées factorielles des lignes et des colonnes obtenues à l'aide de PCAmix, d'autres sorties numériques existent afin de faciliter l'interprétation des résultats. C'est le cas des contributions (des observations et des variables) et des cosinus carrés qui quantifient la qualité de représentation des points (observations et variables) sur les axes factoriels. Nous allons voir dans cette section comment les calculer. De plus, une fois chaque sortie définie, nous présentons un extrait de code R permettant d'obtenir les valeurs numériques à l'aide du package `PCAmixdata` et de la fonction `PCAmix`. Dans la suite, on suppose que le résultat de la fonction `PCAmix` est stocké dans un objet appelé `res.pcamix`. On rappelle les notations suivantes : $f_{ik}$ est la coordonnée factorielle de l'observation $i$ sur l'axe factoriel $k$ et $a^\star_{jk}$ (resp.$a^\star_{sk}$) est la coordonnée factorielle de la variable quantitative $j$ (resp. de la modalité $s$) sur l'axe $k$.

#### 2.3.2.1 Sorties numériques relatives aux observations

**Contributions des observations aux axes factoriels.** La contribution absolue d'une observation $i$ à la variance de l'axe factoriel $k$, notée $c_{ik}$, se calcule de la manière suivante :

$$c_{ik} = \frac{1}{n} {f_{ik}}^2.$$

La contribution relative d'une observation $i$ à un axe factoriel $k$ est égale à sa contribution absolue $c_{ik}$ divisée par la valeur propre $\lambda_k$ associée à l'axe $k$.

Les contributions absolues et relatives des observations s'obtiennent avec les lignes de code suivantes :

```
res.pcamix$ind$contrib
res.pcamix$ind$contrib.pct
```

**Cosinus carrés des observations sur les axes factoriels.** La qualité de représentation d'une observation $i$ sur un axe factoriel $k$ se mesure à l'aide de la valeur appelée cosinus carré et se calcule comme suit :

$$\cos^2(i,k) = \frac{{f_{ik}}^2}{\|d_i\|^2},$$

où $\|d_i\|^2 = \sum_{k=1}^{r} {f_{ik}}^2$ et $r$ (le rang de la matrice $\mathbf{Z}$) est le nombre maximum d'axes factoriels disponibles. Plus une observation est bien projetée sur un axe factoriel, plus son cosinus carré associé sera proche de 1. Inversement si l'observation est mal représentée, son cosinus carré sera proche de 0.
Les valeurs numériques des cosinus carrés des observations s'obtiennent comme suit :

```
res.pcamix$ind$cos2
```

#### 2.3.2.2 Sorties numériques relatives aux variables quantitatives

**Contributions des variables quantitatives aux axes factoriels.** La contribution absolue d'une variable quantitative $j$ à la variance de l'axe factoriel $k$, notée $c_{jk}$, se calcule de la manière suivante :

$$c_{jk} = {a^\star_{jk}}^2.$$

La contribution relative d'une variable quantitative $j$ à un axe factoriel $k$ est égale à sa contribution absolue $c_{jk}$ divisée par la valeur propre $\lambda_k$ associée à l'axe $k$.
Les contributions absolues et relatives des variables quantitatives s'obtiennent avec le code suivant :

```
res.pcamix$quanti$contrib
res.pcamix$quanti$contrib.pct
```

**Cosinus carrés des variables quantitatives sur les axes factoriels.** Le cosinus carré d'une variable quantitative $j$ sur l'axe factoriel $k$, noté $\cos^2(j,k)$ se calcule de la manière suivante :

$$\cos^2(j,k) = \frac{{a^\star_{jk}}^2}{\|d_j\|^2},$$

où $\|d_j\|^2 = \sum_{k=1}^{r} {a^\star_{jk}}^2$ et $r$ est le nombre total d'axes factoriels. Comme pour les observations, plus une variable quantitative est bien projetée sur un axe factoriel, plus son

cosinus carré associé sera proche de 1. Inversement si la variable quantitative est mal représentée, son cosinus carré sera proche de 0.
Les cosinus carrés des variables quantitatives s'obtiennent comme suit :

```
res.pcamix$quanti$cos2
```

#### 2.3.2.3 Sorties numériques relatives aux modalités des variables qualitatives

**Contributions des modalités aux axes factoriels.** La contribution absolue d'une modalité $s$ à la variance de l'axe factoriel $k$, notée $c_{sk}$ se calcule de la manière suivante :

$$c_{sk} = \frac{n_s}{n} {a_{sk}^\star}^2.$$

La contribution relative d'une modalité $s$ à un axe factoriel $k$ est égale à sa contribution absolue $c_{sk}$, divisée par la valeur propre $\lambda_k$ associée à l'axe $k$.
On peut également définir la notion de contribution pour une variable qualitative comme la somme des contributions de ses modalités.
Les contributions absolues et relatives des modalités ainsi que celles des variables qualitatives s'obtiennent avec les trois lignes de code suivantes :

```
res.pcamix$levels$contrib
res.pcamix$levels$contrib.pct
res.pcamix$quali$contrib
```

**Cosinus carrés des modalités sur les axes factoriels.** Le cosinus carré d'une modalité $s$ sur l'axe factoriel $k$, noté $\cos^2(s,k)$, se calcule de la manière suivante :

$$\cos^2(s,k) = \frac{{a_{sk}^\star}^2}{\|d_s\|^2},$$

où $\|d_s\|^2 = \sum_{k=1}^{r} {a_{sk}^\star}^2$. Comme pour les observations et les variables quantitatives, plus une modalité est bien projetée sur un axe factoriel, plus son cosinus carré associé sera proche de 1.
On peut obtenir les cosinus carrés des modalités de la manière suivante :

```
res.pcamix$levels$cos2
```

#### 2.3.2.4 Les "squared loadings" pour étudier le lien entre les variables (quantitatives et qualitatives) et les composantes principales

Nous introduisons ici la notion de "squared loading" entre une variable $\mathbf{x}_j$ (quantitative ou qualitative) et un axe factoriel $\mathbf{f}_k$. Les "squared loadings" permettent de

représenter sur un même graphique les variables quantitatives et les variables qualitatives et ainsi voir leurs liens avec les axes factoriels.
Le "squared loading" entre une variable quantitative $\mathbf{x}_j$ et un axe factoriel (ou composante principale) $\mathbf{f}_k$ est égal à la corrélation au carré, $r^2(\mathbf{x}_j, \mathbf{f}_k)$, entre la variable et la composante principale. Lorsque la variable $\mathbf{x}_j$ est qualitative, son "squared loading" est égal au rapport de corrélation $\eta^2(\mathbf{f}_k|\mathbf{x}_j)$ entre la variable et la composante principale.
Le rapport de corrélation $\eta^2(y|x)$ entre la variable quantitative $y$ et la variable qualitative $x$ est défini comme suit :

$$\eta^2(y|x) = \frac{\sum_{s=1}^{m} n_s(\bar{y}_s - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3.4}$$

où $m$ est le nombre total de modalités de la variable $x$, $n_s$ est le nombre d'observations possédant la modalité $s$, $\bar{y}_s$ est la moyenne de la variable $y$ calculée sur les observations possédant la modalité $s$ et $\bar{y}$ est la moyenne de la variable $y$ calculée sur toutes les observations.
Dans PCAmix, le "squared loading" d'une variable est exactement égal à la contribution de la variable (quantitative ou qualitative) à l'axe factoriel.
Les "squared loadings" de l'ensemble des variables s'obtiennent avec le code suivant :

```
res.pcamix$sqload
```

#### 2.3.2.5 Coefficients des combinaisons linéaires associées aux composantes principales

Nous avons vu que la méthode PCAmix permet d'obtenir des nouvelles variables quantitatives appelées composantes principales qui expliquent la plus grande part d'inertie possible de la matrice $\mathbf{Z}$. Les composantes principales (colonnes de $\mathbf{F}$) sont alors des combinaisons linéaires non corrélées des colonnes de $\mathbf{Z}$ avec :

- une variance maximum : $\lambda_k = \|\mathbf{f}_k\|_{\mathbf{N}} = \mathrm{Var}(\mathbf{f}_k)$,
- un lien maximum avec les variables d'origine :

$$\lambda_k = \sum_{j=1}^{p_1} r^2(\mathbf{x}_j, \mathbf{f}_k) + \sum_{j=p_1+1}^{p_2} \eta^2(\mathbf{f}_k|\mathbf{x}_j). \tag{2.3.5}$$

Les coefficients de ces combinaisons linéaires peuvent être utilisés, par exemple, pour prédire les coordonnées de nouvelles observations sur les composantes principales de PCAmix.
La $k$-ème composante principale peut s'écrire comme une combinaison linéaire des

vecteurs $\mathbf{z}_1, \ldots, \mathbf{z}_{p_1+m}$ (colonnes de $\mathbf{Z}$) pour :

$$\mathbf{f}_k = \mathbf{ZMv}_k = \sum_{j=1}^{p_1} v_{jk}\mathbf{z}_j + \sum_{s=p_1+1}^{p_1+m} \frac{n}{n_s} v_{js}\mathbf{z}_s.$$

On peut montrer facilement que $\mathbf{f}_k$ s'écrit :

$$\mathbf{f}_k = \beta_0 + \sum_{j=1}^{p_1} \beta_j \mathbf{x}_j + \sum_{s=p_1+1}^{p_1+m} \beta_s \mathbf{x}_s$$

où les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_{p_1+m}$ sont les colonnes de $\mathbf{X} = (\mathbf{X}_1|\mathbf{G})$. Cela donne :

$$\begin{aligned}
\beta_0 &= -\sum_{j=1}^{p_1} v_{jk}\frac{\bar{\mathbf{x}}_j}{\sigma_j} - \sum_{s=p_1+1}^{p_1+m} v_{sk}, \\
\beta_j &= v_{jk}\frac{1}{\sigma_j}, \text{ for } k = 1, \ldots, p_1, \\
\beta_s &= v_{sk}\frac{n}{n_s}, \text{ for } s = p_1+1, \ldots, p_1+m,
\end{aligned}$$

avec $\bar{\mathbf{x}}_j$ et $\sigma_j$ respectivement la moyenne empirique et l'écart-type empirique de la colonne $\mathbf{x}_j$.

## 2.4 Illustration de la méthode PCAmix à l'aide du package PCAmixdata

Nous illustrons dans cette section la méthode PCAmix présente dans le package R `PCAmixdata`. Le package contient quatre jeux de données décrivant les mêmes observations. Ces quatre `dataframe` sont rassemblés dans l'objet `gironde` de type "`list`". Nous allons appliquer la méthode PCAmix sur le jeu de données `housing` où $n = 542$ communes de Gironde sont décrites par $p_1 = 3$ variables quantitatives et $p_2 = 2$ variables qualitatives ayant un nombre total de $m = 4$ modalités. Une description des variables utilisées est donnée en Annexe B. L'extrait de code ci-dessous détaille comment charger les données et comment utiliser la fonction `PCAmix`.

```
## chargement du package
library(PCAmixdata)
## chargement de la liste gironde contenant differents dataframe
data(gironde)
## on recupere le jeu de donnees housing
housing<-gironde$housing[-c(1:10), ]
## on lance la methode PCAmix
res.pcamix<-PCAmix(data=housing, rename.level=TRUE, graph=FALSE)
```

La fonction `plot.PCAmix` permet d'afficher différentes sorties graphiques nécessaires à l'interprétation des résultats. Nous détaillons ici l'interprétation des quatre graphiques représentés à la Figure 2.1 au travers de quatre paragraphes. A la fin de chaque paragraphe, on trouve un extrait de code R utilisé pour l'affichage du graphique correspondant.

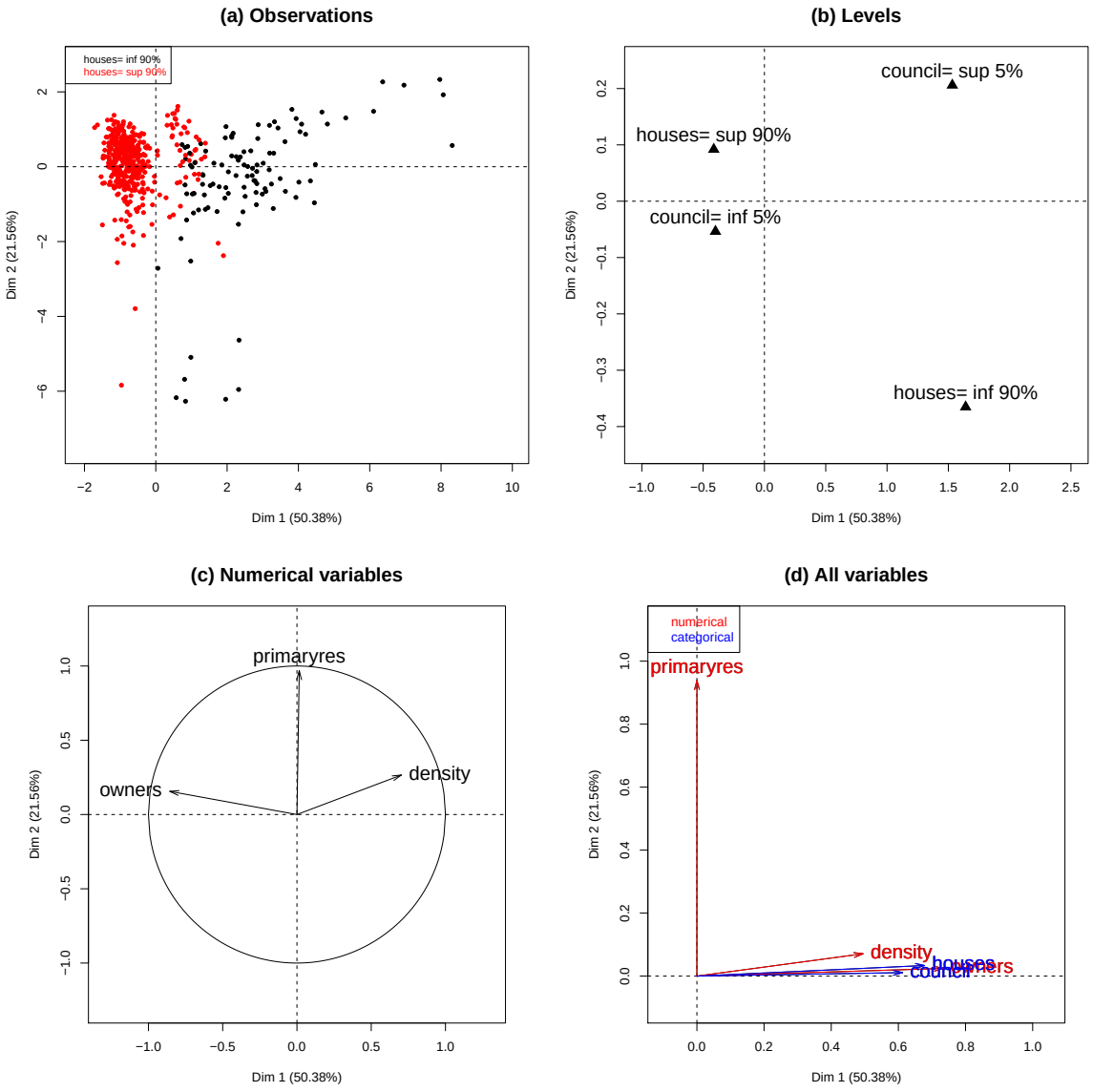



FIGURE 2.1 – (a) Coordonnées factorielles des observations. (b) Coordonnées factorielles des modalités. (c) Cercle des corrélations. (d) "Squared loadings" de toutes les variables.

**Valeurs propres et pourcentage d'inertie expliquée.** Afin de choisir le nombre d'axes factoriels à interpréter, on observe les valeurs propres et les pourcentages d'iner-

tie expliquée associés à l'aide de la commande suivante :

```
res$eig

#       Eigenvalue Proportion Cumulative
#dim 1   2.5268771  50.537541   50.53754
#dim 2   1.0692777  21.385553   71.92309
#dim 3   0.6303253  12.606505   84.52960
#dim 4   0.4230216   8.460432   92.99003
#dim 5   0.3504984   7.009968  100.00000
```

Dans cet exemple, on choisit d'interpréter les résultats sur les deux premières composantes principales (axes factoriels) qui expliquent 71,9% de la variance des données.

**Coordonnées factorielles des observations.** La Figure 2.1(a) représente les coordonnées factorielles des observations sur le plan engendré par les axes factoriels 1 et 2. Chaque observation est colorée en fonction de la modalité de la variable `houses` : "moins de 90% de maisons sur la commune" ou "plus de 90%". On remarque que l'axe 1 est discriminant pour cette variable qualitative. En effet, on retrouve sur la gauche (faibles valeurs de la première composante principale) les communes avec une forte proportion de maisons, alors que sur la droite du graphique (valeurs élevées de la première composante principale) on retrouve les communes avec une faible proportion de maison.

```
plot(res.pcamix, choice="ind", axes=c(1,2), coloring.ind=housing$
   houses, label=FALSE, main="(a) Observations", cex.main=1.5)
```

**Coordonnées factorielles des modalités.** La Figure 2.1(b) représente la carte des modalités des variables qualitatives sur le plan factoriel 1-2, on observe que les communes ayant une forte proportion de maisons sont également celles qui ont un faible pourcentage de logements sociaux (variable `council`).

```
plot(res.pcamix, choice="levels", axes=c(1,2), cex=1.5, xlim=c(-1,2.5)
   , main="(b) Levels", cex.main=1.5)
```

**Cercle des corrélations des variables quantitatives.** La représentation des coordonnées factorielles des variables quantitatives (appelée cercle des corrélations) sur la Figure 2.1(c) indique que la densité de population (variable `density`) est corrélée négativement au nombre de propriétaires de leur logement (variable `owners`). De plus ces deux variables sont fortement corrélées à la première composante principale (Axe 1).

```
plot(res.pcamix, choice="cor", axes=c(1,2), cex=1.5, main="(c)
   Numerical variables", cex.main=1.5)
```

**"Squared loadings" de toutes les variables.** La Figure 2.1(d) représente sur un même graphique les variables quantitatives et qualitatives en fonction de leurs "squared loadings". On observe que toutes les variables sont liées à la première composante, exceptée la variable indiquant le nombre de résidences principales sur la commune (variable `primaryres`) qui est fortement corrélée à la seconde composante principale et donc non corrélée aux autres variables.

```
plot(res.pcamix, choice="sqload", axes=c(1,2), cex=1.5, leg.sqload=
   TRUE, xlim=c(-0.1,1.05), main="(d) All variables", cex.main=1.5)
```

**Prédiction des coordonnées d'observations supplémentaires.** La fonction `predict.PCAmix` permet de calculer les coordonnées factorielles de nouvelles observations n'ayant pas été incluses dans l'analyse. Nous allons voir comment l'utiliser dans l'extrait de code ci-dessous :

```
## on recupere 10 nouvelles observations non incluses dans l'analyse
newind<-gironde$housing[1:10, ]
## on utilise la fonction predict pour calculer les coordonnees
   factorielles des nouvelles observations
predict.PCAmix(object=res.pcamix, data.new=newind)
```

## 2.5 Conclusion

Ce chapitre a permis de présenter la méthode PCAmix en utilisant une approche basée sur une GSVD. L'ensemble des sorties numériques relatives aux différents "objets" (variables quantitatives, variables qualitatives et leurs modalités, observations) ont été décrites. Nous avons également détaillé l'utilisation de la méthode au sein du package `PCAmixdata`. Un exemple simple sur des données réelles a permis de présenter des extraits de code afin d'obtenir les résultats graphiques et numériques nécessaires à l'interprétation des résultats.
La méthode PCAmix ne sera pas utilisée en tant que telle pour la construction d'indicateurs composites de la qualité de vie. Cependant nous allons voir que cette méthode est au cœur du chapitre suivant sur la classification de variables, mais également au sein du chapitre présentant l'analyse factorielle multiple de données mixtes (MFAmix).

# 3 Classification de variables : la méthode hclustvar

## Sommaire



## 3.1 Introduction

La classification de variables (aussi appelée clustering de variables) est une alternative intéressante aux méthodes classiques d'analyses factorielles pour la réduction de dimension et donc pour la construction d'indicateurs composites de qualité de vie. A notre connaissance, l'approche par classification de variables n'a jamais été utilisée pour la construction d'indicateurs de qualité de vie. Le but du clustering de variables est de rassembler dans un même cluster les variables qui se ressemblent, c'est à dire qui

apportent la même information, afin de construire des clusters homogènes de variables. Afin d'éviter toute confusion nous parlerons de "cluster" de variables et lorsque nous ferons de la classification d'observations, nous parlerons de "classes" d'observations.
L'approche la plus couramment utilisée en clustering de variables consiste à calculer une matrice contenant des mesures d'association entre variables puis d'appliquer une méthode de classification d'observations sur cette matrice. Pour les variables quantitatives une mesure d'association fréquemment utilisée est le coefficient de corrélation. Pour les variables qualitatives, on peut calculer des mesures d'associations, comme la mesure du chi-deux par exemple. Il arrive également que pour effectuer du clustering de variables certaines personnes utilisent la distance euclidienne entre les variables comme mesure d'association. Cependant cette mesure est pertinente pour calculer des distances entre observations mais elle n'a que peu de sens pour des calculs de distances entre variables.
Il existe également d'autres méthodes spécialement dédiées à la classification de variables. Pour les variables quantitatives, la méthode la plus connue est la procédure `VARCLUS` du logiciel SAS [1]. Deux autres méthodes basées sur l'analyse en composantes principales (ACP) existent pour la classification de variables quantitatives. La méthode "Clustering of variables around Latent Variables" (CLV), voir par exemple Vigneau and Qannari (2003) et Vigneau et al. (2006), est implémentée dans le package `ClustVarLV`, voir Vigneau and Chen (2015). La seconde méthode, basée également sur l'ACP est la méthode appelée "Diametrical clustering method" développée par Dhillon et al. (2003). A notre connaissance la classification de variables qualitatives (ou d'un mélange de variables quantitatives et de variables qualitatives) a été moins étudiée.
Le package `ClustOfVar`, voir Chavent et al. (2015), contient deux méthodes de classification de variables quantitatives et qualitatives. La méthode hclustvar est basée sur un algorithme de clustering ascendant hiérarchique et la méthode kmeansvar est une méthode de type k-means. Ces deux méthodes sont basées sur des algorithmes différents mais elles optimisent le même critère d'homogénéité de partition qui sera présenté dans ce chapitre, voir également Chavent et al. (2012). De plus, un des avantages de ces deux méthodes pour la construction d'indicateurs composites est qu'elles associent à chaque cluster de variables, une variable synthétique qui est la plus liée possible aux variables du cluster. Cette variable synthétique peut ainsi jouer le rôle d'indicateur composite de qualité de vie. De plus, quel que soit le type de variables au sein d'un cluster, la variable synthétique associée est toujours quantitative. Ces variables synthétiques sont assez simples à interpréter car, par construction, elles s'écrivent comme une combinaison linéaire des variables quantitatives et des modalités des variables qualitatives du cluster et non pas de toutes les variables comme par exemple les composantes princi-

1. SAS Institute Inc. 2015. SAS/STAT® 14.1 User's Guide

pales de PCAmix. De plus, aucune condition d'orthogonalité n'est imposée entre les différentes variables synthétiques associées aux différents clusters.
La Section 3.2 présente le critère d'homogénéité de partition défini dans le package `ClustOfVar` ainsi que l'algorithme de classification ascendante hiérarchique de la méthode hclustvar. Un exemple simple sera donné afin de se familiariser avec les fonctions du package et leur utilisation. La Section 3.3 présente l'utilisation de la méthode hclustvar dans le but de construire des indicateurs composites de qualité de vie à l'échelle communale.

## 3.2 Le package `ClustOfVar`

Le package `ClustOfVar` contient deux méthodes de classification de variables. La méthode hclustvar utilise une approche de type classification ascendante hiérarchique alors que la méthode kmeansvar utilise une approche de type k-means. Cependant, ces deux méthodes optimisent un seul et même critère d'homogénéité de partition. Nous présentons dans cette section le critère d'homogénéité utilisé dans le package `ClustOfVar` au sein des fonctions `hclustvar` et `kmeansvar`. Puis, nous définissons la variable synthétique $\mathbf{z}_k$ associée aux différents clusters de variables. Par la suite, nous détaillons l'algorithme de la méthode hclustvar.
Nous introduisons les notations suivantes. Soit $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p_1}\}$ un ensemble de $p_1$ variables quantitatives et $\{\mathbf{y}_1, \ldots, \mathbf{y}_{p_2}\}$ un ensemble de $p_2$ variables qualitatives. Et $\mathbf{X}$ (resp. $\mathbf{Y}$) la matrice $n \times p_1$ (resp. $n \times p_2$) contenant les variables quantitatives (resp. les variables qualitatives), avec $n$ le nombre d'observations. On notera $\mathbf{x}_j \in \mathbb{R}^n$ la $j$-ème colonne de $\mathbf{X}$ et $\mathbf{y}_j \in \mathcal{M}_j^n$ la $j$-ème colonne de $\mathbf{Y}$, avec $\mathcal{M}_j$ l'ensemble des modalités de $\mathbf{y}_j$. Pour finir, on note $P_K = (C_1, \ldots, C_K)$ la partition des $p = p_1 + p_2$ variables en $K$ clusters.

### 3.2.1 Critère d'homogénéité d'un cluster de variables

On note $\mathcal{H}(P_K)$ le critère d'homogénéité de la partition $P_K$. Ce critère est défini par :

$$\mathcal{H}(P_K) = \sum_{k=1}^{K} H(C_k), \tag{3.2.1}$$

où $H(C_k)$ est le critère d'homogénéité du cluster $C_k$. Le critère d'homogénéité d'un cluster est défini comme la somme des liaisons entre les variables du cluster et la variable synthétique du cluster notée $\mathbf{z}_k$. Cette variable synthétique est un vecteur de $\mathbb{R}^n$ qui résume les variables du cluster. Elle sera donc par nature toujours quantitative.

Le critère d'homogénéité de cluster est défini comme suit :

$$H(C_k) = \sum_{\mathbf{x}_j \in C_k} r^2(\mathbf{x}_j, \mathbf{z}_k) + \sum_{\mathbf{y}_j \in C_k} \eta^2(\mathbf{y}_j|\mathbf{z}_k), \tag{3.2.2}$$

Le premier terme $r^2(\mathbf{x}_j, \mathbf{z}_k)$ est la corrélation de Pearson au carré entre la variable quantitative $\mathbf{x}_j$ et la variable synthétique $\mathbf{z}_k$. Le second terme $\eta^2(\mathbf{y}_j|\mathbf{z}_k) \in [0, 1]$, défini à l'équation (2.3.4), désigne le rapport de corrélation mesurant la part de la variance de $\mathbf{z}_k$ expliquée par les modalités de la variable qualitative $\mathbf{y}_j$.
L'homogénéité d'un cluster est maximale quand toutes les variables quantitatives sont corrélées (positivement ou négativement) à la variable synthétique $\mathbf{z}_k$ et quand tous les rapports de corrélation des variables qualitatives sont égaux à 1. Cela implique que toutes les variables du cluster $C_k$ sont fortement liées entre elles.

### 3.2.2 Définition de la variable synthétique $\mathbf{z}_k$

La variable synthétique $\mathbf{z}_k$ est définie comme le vecteur de $\mathbb{R}^n$ qui maximise le critère d'homogénéité défini à l'équation (3.2.2). C'est la variable la plus liée, au sens du critère d'homogénéité, à l'ensemble des variables du cluster $C_k$. Elle est définie comme suit :

$$\mathbf{z}_k = \arg\max_{\mathbf{u} \in \mathbb{R}^n} \left\{ \sum_{\mathbf{x}_j \in C_k} r^2(\mathbf{x}_j, \mathbf{u}) + \sum_{\mathbf{y}_j \in C_k} \eta^2(\mathbf{y}_j|\mathbf{u}) \right\}.$$

On a les propriétés suivantes :

- $\mathbf{z}_k$ est la première composante principale issue de PCAmix appliquée à la matrice $[\mathbf{X}_k, \mathbf{Y}_k]$ où $\mathbf{X}_k$ (resp. $\mathbf{Y}_k$) est la matrice formée par les colonnes de $\mathbf{X}$ (resp. $\mathbf{Y}$) appartenant au cluster $C_k$.
- Comme nous l'avons vu à l'équation (2.3.5), on a :

$$H(C_k) = \lambda_k^1, \tag{3.2.3}$$

  où $\lambda_k^1 = Var(\mathbf{z}_k)$ est la première valeur propre issue de PCAmix appliquée aux variables du groupe $C_k$.

A partir des équations (3.2.1) et (3.2.3) on réécrit le critère d'homogénéité de partition de la manière suivante :

$$\mathcal{H}(P_k) = \sum_{k=1}^{K} \lambda_k^1. \tag{3.2.4}$$

Nous allons voir par la suite comment définir une mesure d'agrégation entre clusters et comment mettre en œuvre l'algorithme de la méthode de classification ascendante hiérarchique hclustvar.

### 3.2.3 La méthode hclustvar

L'objectif est de trouver une partition qui maximise le critère d'homogénéité $\mathcal{H}(P_K)$. Pour cela, on propose d'utiliser un algorithme de classification ascendante hiérarchique, qui construit un ensemble de $p$ partitions de variables emboitées. A chaque étape de l'algorithme, on obtient la meilleure partition, au sens du critère $\mathcal{H}(P_K)$, parmi toutes les partitions existantes issues du rassemblement de deux clusters de la partition obtenue à l'étape précédente. L'algorithme fonctionne de la manière suivante :

**Etape $s = 0$ :** On commence avec la partition des $p$ singletons. Chaque variable appartient à un cluster différent.

**Etape $s = 1, \ldots, p-1$ :** On agrège les deux clusters $C_l$ et $C_m$ de la partition en $p-s+1$ clusters pour obtenir une partition en $p-s$ clusters. On agrège les deux clusters tels que la perte de l'homogénéité de partition issue du rassemblement soit la plus petite possible. Ainsi, on agrège les deux clusters ayant la mesure d'agrégation $\delta(C_l, C_m)$, définie ci-dessous la plus petite.

$$\begin{aligned}\delta(C_l, C_m) &= \mathcal{H}(P_{p-s+1}) - \mathcal{H}(P_{p-s}) \\ &= \lambda_1^1 + \cdots + \lambda_l^1 + \lambda_m^1 + \cdots + \lambda_K^1 - (\lambda_1^1 + \cdots + \lambda_{(l\cup m)}^1 + \cdots + \lambda_K^1) \\ &= \lambda_l^1 + \lambda_m^1 - \lambda_{(l\cup m)}^1 \end{aligned} \tag{3.2.5}$$

où $\lambda_{(l\cup m)}^1$ est la première valeur propre issue de PCAmix appliquée sur le cluster $C_l \cup C_m$.

**Etape $p$ :** La partition en un seul cluster contenant toutes les variables est obtenue.

La hauteur d'une classe $C = C_l \cup C_m$ dans le dendrogramme est définie par $h(C) = \delta(C_l, C_m)$. On a bien $h(C) \geqslant 0$ mais la propriété de croissance monotone de la mesure d'agrégation, "$C_l \subset C_m \Rightarrow h(C_l) \leqslant h(C_m)$" n'est pas démontrée. Notons qu'en pratique, aucune inversion n'a été observée sur les différents jeux de données (réelles ou simulées) utilisés.

### 3.2.4 Illustration de hclustvar sur un exemple

Le package `ClustOfVar` est illustré ici sur les jeux de données de l'objet `gironde` présent dans le package `PCAmixdata`. L'objet `gironde` est un objet de type "`list`" contenant quatre `dataframe` relatifs à quatre groupes de variables décrivant les mêmes observations (ici les communes du département de la Gironde). Les quatre groupes de variables représentent quatre thématiques relatives aux conditions de vie : l'emploi (groupe quantitatif `employment`, le logement (groupe mixte `housing`), l'accès à différents services (groupe qualitatif `services`) et l'environnement (groupe quantitatif `environment`). La description des quatre jeux de données est donnée en Annexe B.

Nous allons rassembler ici ces quatre dataframe en une seule matrice sur laquelle nous appliquerons la fonction hclustvar. Dans un premier temps, nous verrons comment obtenir une partition de variables, puis dans un second temps nous détaillerons les outils disponibles pour interpréter cette partition.

**Chargement des données et classification hiérarchique avec hclustvar.** L'extrait de code ci-dessous permet de charger les données gironde contenues dans le package PCAmixdata puis d'exécuter la méthode hclustvar.

```
## chargement de la liste gironde contenant les differents dataframe
## dans le package PCAmixdata
library(PCAmixdata)
data(gironde)
## rassemblement des quatre datataframe dans un seul
dat<-cbind(gironde$employment[1:200, ], gironde$housing[1:200, ],
           gironde$services[1:200, ], gironde$environment[1:200, ])
## on lance la methode hclustvar
res<-hclustvar(X.quanti=X.qt, X.quali=X.ql)
```

**Le dendrogramme.** Le code R suivant permet d'obtenir le dendrogramme représenté à la Figure 3.1 :

```
## plot du dendrogramme
plot(res, type="tree")
```

L'analyse du dendrogramme peut être utile pour choisir un nombre $K$ de clusters pertinent.

**La partition en 3 clusters.** La fonction hclustvar renvoie une hiérarchie indicée des variables. Pour obtenir une partition en $K$ clusters, il est nécessaire de "couper" cette hiérarchie. Ici, on choisit arbitrairement de retenir la partition en $K = 3$ clusters. Le découpage de la hiérarchie est effectué grâce à la fonction cutreevar à l'aide du code suivant :

```
## obtention de la typologie en 3 clusters
cut3<-cutreevar(res, k=3)
```

**Sorties numériques.** L'objet cut3 créé grâce à la fonction cutreevar est un objet de classe ''clustvar''. Bien que la fonction kmeansvar ne soit pas présentée ici, on notera que le résultat de cette fonction est également un objet de classe ''clustvar''. Différentes sorties numériques sont disponibles au sein de cette classe d'objet. Ces sorties peuvent aider à l'interprétation des clusters de variables et des variables synthétiques qui leurs sont associées. Les sorties numériques disponibles sont les suivantes :

— L'objet cut3$cluster est un vecteur d'entiers $\in \{1, K\}$ de taille $p$ indiquant l'appartenance de chaque variable aux différents clusters.

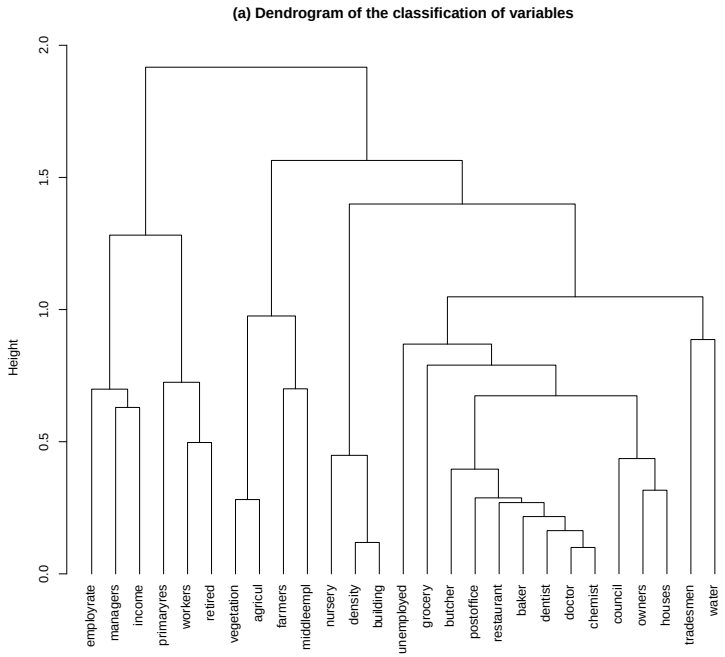



FIGURE 3.1 – Dendrogramme de la hiérarchie des 27 variables des données `gironde`.

```
cut3$cluster[1:5]
# farmers  tradesmen   managers    workers unemployed
#       1          2          3          3          2
```

Ainsi, on voit par exemple que la variable `farmers` est dans le cluster 1 et les variables `tradesmen` et `unemployed` sont dans le cluster 2.

— L'objet `cut3$size` contient le nombre de variables de chaque cluster.

```
cut3$size
# cluster1 cluster2 cluster3
#        4       17        6
```

— L'objet `cut3$wss` contient l'homogénéité de chaque cluster, telle qu'elle est définie à l'équation (3.2.2).

```
round(cut3$wss, 2)
#cluster1 cluster2 cluster3
#     2.04     8.58     2.17
```

Cependant, l'homogénéité du cluster étant liée au nombre de variables du cluster, cette valeur n'est pas toujours très parlante. Une valeur plus explicite est le pourcentage d'inertie expliquée (inertie apportée par les variables du cluster) par la variable synthétique. Ce pourcentage d'inertie expliquée est égal à l'homogénéité du cluster divisée par l'inertie totale du cluster qui vaut : $p_1^{(k)} + m^{(k)} - p_2^{(k)}$, où $p_1^{(k)}$ (resp. $p_2^{(k)}$) est le nombre de variables quantitatives (resp.qualitatives)

du cluster $C_k$ et $m^{(k)}$ est le nombre total de modalités des variables qualitatives du cluster $C_k$.

— L'objet `cut3$scores` contient les coordonnées factorielles (ou scores) des $n$ observations sur chacune des $K = 3$ variables synthétiques. Par construction ces variables sont centrées et leur variance est égale à la première valeur propre de PCAmix appliquée sur le cluster associé.

```
head(cut3$scores)

#                       cluster1  cluster2   cluster3
#ABZAC                 0.4365725 -1.190939 -0.4216443
#AILLAS               -0.3396515  1.040416 -1.1690794
#AMBARES-ET-LAGRAVE -0.2467724 -5.639646  0.9820745
#AMBES                 0.2236974 -2.407998 -0.2960775
#ANDERNOS-LES-BAINS -2.1027407 -4.401294 -1.7624083
#ANGLADE               1.7683769  1.394320 -1.2140944
```

— L'objet `cut3$coef` contient les coefficients de la combinaison linéaire des variables des différents clusters servant à calculer les scores des observations sur les variables synthétiques. Le calcul des coefficients et de la combinaison linéaire est donné à la Section 2.3.2.5. Les coefficients nécessaires à l'obtention des scores sur la première variable synthétique sont obtenus comme suit :

```
cut3$coef$cluster1
#                    [,1]
#const       0.22256157
#farmers     0.09633027
#middleempl -0.06997107
#vegetation -0.02502510
#agricul     0.02558356
```

— L'objet `cut3$var` est une liste de taille $K$ contenant pour chaque cluster les squared loadings (corrélation au carré pour les variables quantitatives et rapport de corrélation pour les variables qualitatives) entre les variables du cluster et la variable synthétique associée. De plus, pour les variables quantitatives la corrélation avec la variable synthétique est également donnée. Cela permet de connaître le sens de la liaison. Cette sortie est utile afin de donner un sens aux différentes variables synthétiques. Pour cela, on regarde quelles variables du cluster sont le plus liées à la variable synthétique associée. Cette sortie numérique s'obtient de la manière suivante pour le cluster 1 :

```
cut3$var$cluster1
#           squared loading                  corre (numeric
   variables)
#agricul          0.8054386                          0.8974623
#vegetation       0.6157969                         -0.7847273
#farmers          0.3993664                          0.6319544
#middleempl       0.2222496                         -0.4714336
```

Les valeurs numériques disponibles pour les 3 clusters sont rassemblées dans la Table 3.1.

| Cluster | Variables | Squared loading | Corrélations pour les variables quantitatives |
|---|---|---|---|
| Cluster 1 | agricul | 0.81 | 0.90 |
| | vegetation | 0.62 | -0.78 |
| | farmers | 0.40 | 0.63 |
| | middleempl | 0.22 | -0.47 |
| Cluster 2 | chemist | 0.86 | - |
| | doctor | 0.78 | - |
| | dentist | 0.76 | - |
| | baker | 0.74 | - |
| | restaurant | 0.70 | - |
| | butcher | 0.66 | - |
| | postoffice | 0.63 | - |
| | houses | 0.63 | - |
| | owners | 0.57 | 0.75 |
| | council | 0.55 | - |
| | density | 0.45 | -0.67 |
| | building | 0.42 | -0.65 |
| | nursery | 0.38 | - |
| | grocery | 0.23 | - |
| | unemployed | 0.15 | -0.39 |
| | water | 0.04 | -0.21 |
| | tradesmen | 0.03 | 0.16 |
| Cluster 3 | retired | 0.58 | -0.76 |
| | employrate | 0.48 | 0.70 |
| | primaryres | 0.41 | 0.64 |
| | managers | 0.34 | 0.58 |
| | income | 0.31 | 0.55 |
| | workers | 0.04 | 0.20 |

TABLE 3.1 – "Squared loadings" entre les variables de chaque cluster et la variable synthétique associée.

**Sorties graphiques.** La fonction `plot` appliquée à un objet de classe `"clustvar"` permet de représenter sur un graphique une partie des résultats contenus dans l'objet `cut3$var`. En effet lorsque le cluster contient des variables quantitatives, la corrélation entre ces variables et la variable synthétique du cluster est représentée afin d'avoir une vision rapide des liaisons entres les variables quantitatives et la variable synthétique. Lorsque le cluster contient des variables qualitatives, on obtient également un graphique qui représente les coordonnées factorielles des modalités des variables qualitatives sur la variable synthétique du cluster qui est, on le rappelle, la première composante principale de PCAmix appliquée sur le cluster. Cette représentation permet dans un premier temps de visualiser les modalités qui s'associent (s'attirent ou se repoussent) dans les clusters. De plus, grâce à la relation barycentrique de l'analyse des correspondances multiples (ACM), on peut interpréter les scores des observations (ici les communes)

sur les variables synthétiques en fonction des modalités. Les graphiques associés aux différents clusters sont représentés à la Figure 3.2. On les obtient grâce au code suivant :

```
plot(cut.3)
```

**Interprétation des variables synthétiques.** L'examen des graphiques de la Figure 3.2 et des valeurs numériques contenues dans la Table 3.1 permet de donner du sens aux trois variables synthétiques :

- La variable synthétique 1 est associée à un cluster contenant 4 variables quantitatives. Elle est corrélée positivement aux variables `agricul` (0.90) et `farmers` (0.63). Alors qu'elle est corrélée négativement aux variables `vegetation` (-0.78) et `middleempl`(-0.47). Ainsi, une commune ayant une forte valeur sur cette variable synthétique est une commune avec une forte proportion de territoires et d'emplois agricoles.
- La variable synthétique 2 est associée à un cluster contenant 11 variables quantitatives et 6 variables qualitatives. Elle est corrélée positivement à la variable quantitative `owners` et négativement aux variables `density`, `building` et dans une moindre mesure aux variables `unemployed` et `water`. On remarque également que le rapport de corrélation entre certaines variables qualitatives et la variable synthétique est assez élevé, c'est le cas par exemple des variables qualitatives `chemist`, `doctor` ou `dentist`. Afin de comprendre quelles modalités sont associées aux différentes valeurs de la variable synthétique, il est nécessaire de regarder les coordonnées factorielles des modalités sur la variable synthétique, qui est également la première composante principale issue de PCAmix appliquée au cluster. Ces valeurs sont représentées sur la Figure 3.2(d). L'examen de cette figure montre clairement un gradient relatif à la présence de services sur la commune. En effet des valeurs négatives de la variable synthétique 2 sont associées à une absence de services (`doctor=0`, `baker=0`,...). Alors que des valeurs positives de la variable synthétique indiquent une présence de services sur la commune (`nursery=1 or +`, `chemist=2 or +`,etc). Cela veut dire qu'une commune ayant une forte valeur sur cette variable synthétique est une commune urbaine avec une forte proportion de bâtiments, une densité de population élevée ainsi qu'une offre importante de services.
- La variable synthétique 3 est associée à un cluster contenant 6 variables quantitatives. Elle est corrélée positivement au variables `employrate`, `primaryres`, `managers` et `income`, elle est corrélée négativement à la variable `retired`. Ainsi, une commune possédant une forte valeur sur cette variable synthétique est une commune avec un taux d'emploi élevé et un nombre important de résidences principales.

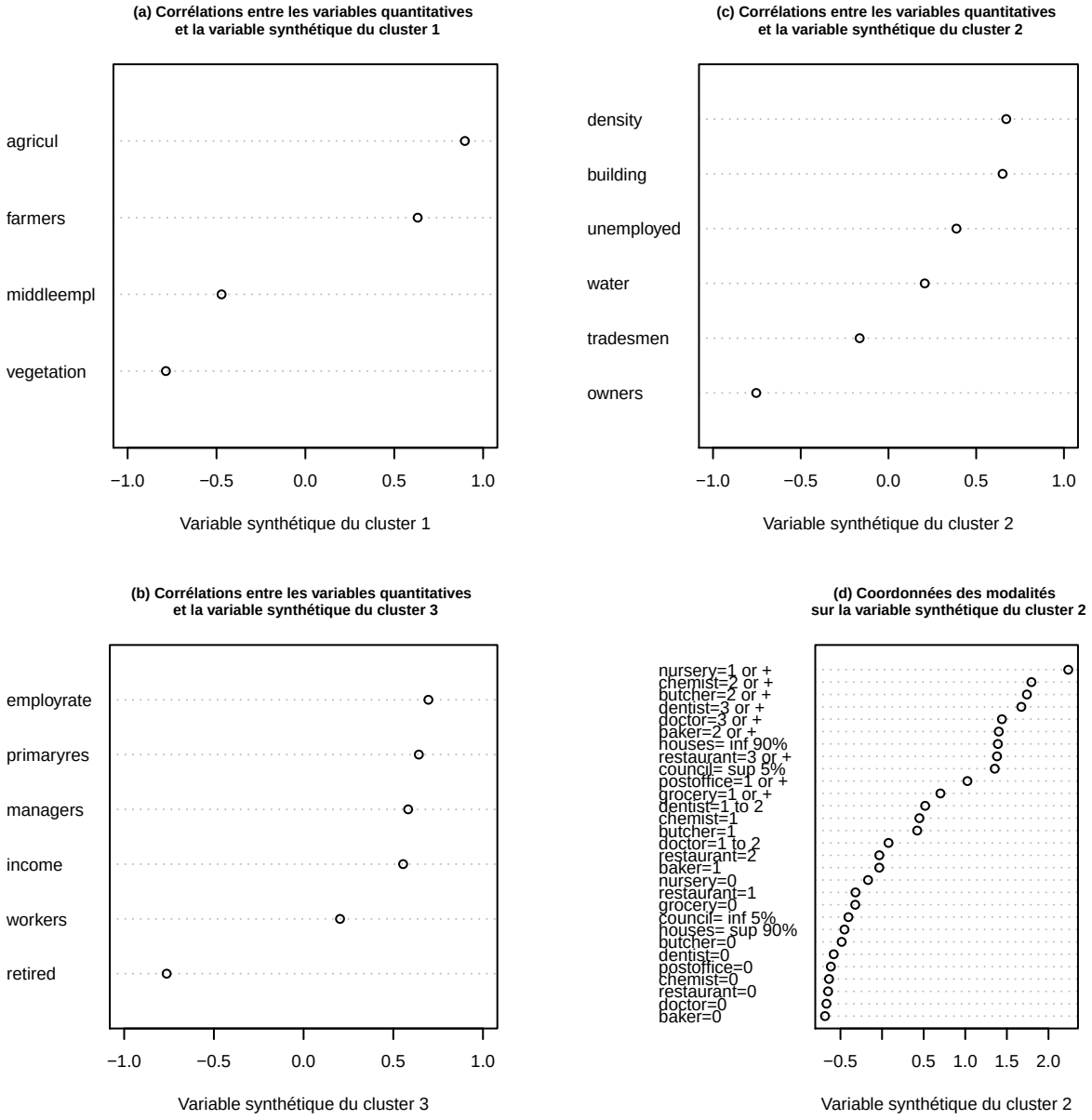



FIGURE 3.2 – (a, b, c) Corrélations entre les variables quantitatives et les variables synthétiques des clusters 1 à 3. (d) Coordonnées factorielles des modalités des variables qualitatives sur la variable synthétique du cluster 2.

| Variable synthétique | Valeurs négatives | Valeurs positives |
|---|---|---|
| VS 1 : Environnement | Commune végétalisée<br>Peu de territoires agricoles<br>Peu d'emplois agricoles | Commune agricole<br>Peu de territoires végétalisés<br>Pourcentage d'emplois agricoles élevé |
| VS 2 : Services et urbanisation | Peu de services<br>Proportion importante de propriétaires<br>Faible densité de population<br>Peu de bâtiments | Beaucoup de services<br>Peu de propriétaires<br>Forte densité de population<br>Forte proportion de bâtiments |
| VS 3 : Emploi | Proportion importante de retraités<br>Faible taux d'emploi<br>Peu de résidences principales | Peu de retraités<br>Taux d'emploi élevé<br>Nombre important de résidences principales |

TABLE 3.2 – Lecture des variables synthétiques.

Cette interprétation des variables synthétiques des clusters permet de leur donner un nom en fonction des variables qui leur sont le plus corrélées. On peut également relier les valeurs d'une variable synthétique (positives ou négatives) aux valeurs des variables du cluster. Cela permet de connaître le profil d'une commune en fonction de son score sur les différentes variables synthétiques. On crée ainsi la Table 3.2 qui est une aide à la lecture des variables synthétiques. On peut par exemple connaître le profil de la commune d'Andernos les Bains grâce à ces scores sur les différentes variables synthétiques. Les scores sont obtenus à l'aide du code R suivant :

```
res.3$scores["ANDERNOS-LES-BAINS", ]
# cluster1  cluster2  cluster3
#    -2.10     -4.40     -1.76
```

Ainsi, la commune d'Andernos est une commune relativement végétalisée (faible valeur sur la variable synthétique associée au cluster 1). Elle possède une faible densité de population, peu de bâtiments et peu de services (valeur négative sur la deuxième variable synthétique). C'est également une commune avec une forte proportion de retraités, un taux d'emploi assez faible et peu de résidences principales (valeur négative sur la troisième variable synthétique). Ce tableau de lecture des variables synthétiques sera également utilisé dans la section suivante afin d'interpréter les différentes classes d'une typologie des observations réalisée sur les variables synthétiques.

## 3.3 Construction d'indicateurs composites de qualité de vie à l'aide du package ClustOfVar

Cette section est dédiée à la construction d'indicateurs composites de qualité de vie à l'aide de la méthode de classification de variables hclustvar présentée précédemment.

Cette approche hiérarchique nous permet de voir les associations entre les variables et de "comprendre" la structuration de la qualité de vie. De plus, les variables synthétiques issues de la méthode hclustvar peuvent être considérées comme des indicateurs composites. En effet, nous avons vu que la variable synthétique $\mathbf{z}_k$ est la variable la plus liée, au sens du critère d'homogénéité, à l'ensemble des variables du cluster $C_k$. De plus, la variable synthétique $\mathbf{z}_k$ est la première composante principale issue de PCAmix, elle peut donc s'écrire comme une combinaison linéaire des variables du cluster $C_k$, comme nous l'avons vu à la Section 2.3.2.5. La méthode hclustvar sera appliquée sur une zone d'étude liée à la Garonne sur des variables relatives à la qualité de vie. De plus, la procédure sera réalisée sur le même échantillon de communes sur deux années différentes : 1999 et 2009.
L'objectif de ce chapitre est double. D'une part, nous souhaitons comprendre comment se structurent les différents indicateurs composites de la qualité de vie, c'est à dire savoir quelles variables se ressemblent et se regroupent dans des clusters homogènes. D'autre part, une fois les indicateurs composites construits, nous souhaitons les utiliser pour réaliser une typologie des communes étudiées. Ainsi, nous pourrons voir quelles communes ont le même profil et quelles sont leurs caractéristiques. Le fait de réaliser cette procédure sur deux années différentes avec un intervalle de temps de 10 ans nous permettra d'observer les différences de structuration des indicateurs de qualité de vie mais aussi de mieux comprendre l'évolution des profils de communes.

### 3.3.1 Présentation de la zone d'étude, des données et de la méthodologie adoptée

**La Zone Garonne-Gironde.** L'analyse de la qualité de vie à l'aide de la méthode hclustvar à été menée sur le système socio-écologique de la Zone Garonne-Gironde (ZGG) du sud-ouest de la France. Ces deux rivières se jettent dans l'océan Atlantique au niveau de l'estuaire de la Gironde, un des plus gros estuaires d'Europe. Le périmètre de notre étude englobe environ 3300 communes, pour une population totale d'environ 4.3 millions de personnes. Ces communes sont situées sur une bande de 50km de chaque coté de la Garonne et de l'estuaire de la Gironde. Cela représente une surface d'environ 50000 km$^2$ (environ 10% de la surface totale de la France). Contrairement à d'autres systèmes estuariens européens (Rhin, Seine), l'estuaire de la Gironde est principalement constitué de territoires ruraux avec une faible densité de population. Les zones les plus densément peuplées de la ZGG se trouvent à deux endroits : la métropole toulousaine et la métropole bordelaise. L'analyse de cette zone d'étude liée au fleuve permettra de mieux comprendre les spécificités des communes et leurs évolutions. En effet, l'analyse des changements des conditions de vie à l'échelle des communes est une étape clé pour

la compréhension des processus qui ont contrôlés la qualité de vie de leurs habitants : l'évolution de l'agriculture, l'étalement urbain, le vieillissement de la population. Cette analyse peut permettre de comprendre si ces changements peuvent être considérés comme des facteurs de vulnérabilité.

**Les données utilisées.** Nous avons présenté à la Section 1.2.2 le type de données que nous allons utiliser pour la construction des indicateurs composites. Dans ce chapitre, nous utilisons deux jeux de données relatifs à deux années différentes. Nous avons utilisé 55 variables en 1999. Les variables prises en compte en 2009 sont au nombre de 46 car certaines variables mesurées en 1999 par l'INSEE ne l'étaient plus en 2009. La description et la moyenne de chaque variable utilisée est donnée en Annexe C.

**Méthodologie adoptée.** La méthode hclustvar est utilisée sur les données de 1999 et de 2009. Les variables synthétiques obtenues sont les indicateurs composites de qualité de vie relatifs aux deux années. Dans un premier temps nous expliquons comment choisir un nombre d'indicateurs en s'aidant du dendrogramme issu de la classification ascendante hiérarchique de variables. Puis, nous interpréterons ces indicateurs à l'aide des variables contenues dans chaque cluster. Par la suite, nous réalisons une classification ascendante hiérarchique des communes afin de construire des groupes de communes ayant des caractéristiques semblables sur ces indicateurs de qualité de vie.

### 3.3.2 Résultats de hclustvar et typologie des communes pour l'année 1999

#### 3.3.2.1 Construction des indicateurs composites de l'année 1999 avec hclustvar

La méthode hclustvar a été réalisée sur les données de 1999. La première étape consiste à choisir un nombre de clusters pertinent. Pour cela on peut s'aider du dendrogramme présenté à la Figure 3.3. A chaque étape d'agrégation de l'algorithme, la hauteur du dendrogramme correspond à la mesure d'agrégation entre deux clusters de variables définie à l'équation (3.2.5). Ainsi un saut observé sur le dendrogramme correspond à l'agrégation de deux clusters relativement différents. Cependant le choix du nombre de clusters et donc du nombre d'indicateurs composites (variables synthétiques associées aux différents clusters) n'est pas uniquement basé sur des critères statistiques. En effet le but premier est de pouvoir donner un sens aux différents indicateurs en fonction des variables présentes dans chaque cluster. Ainsi, nous choisissons de couper le dendrogramme en 5 clusters distincts de variables. Les résultats présentés à la Table 3.3 montrent que les 5 clusters sont de tailles différentes. Pour chaque clus-

| | Variables quantitatives | Variables qualitatives | Nombre total de variables | Homogénéité du cluster | Pourcentage d' inertie expliquée |
|---|---|---|---|---|---|
| Cluster 1 | 7 | 0 | 7 | 2.7 | 37.9 |
| Cluster 2 | 14 | 0 | 14 | 4.0 | 28.5 |
| Cluster 3 | 5 | 3 | 8 | 4.1 | 50.8 |
| Cluster 4 | 4 | 0 | 4 | 2.0 | 49.3 |
| Cluster 5 | 4 | 18 | 22 | 10.6 | 37.9 |

TABLE 3.3 – Composition des 5 clusters de variables pour l'année 1999.

ter, nous détaillons sa taille (nombre de variables), son homogénéité telle que décrite à l'équation (3.2.3) ainsi que le pourcentage d'inertie expliquée par la variable synthétique associée. On rappelle que le pourcentage d'inertie expliquée d'un cluster est égal à son homogénéité divisée par la variance totale du cluster, égale à $p_1^{(k)} + m^{(k)} - p_2^{(k)}$. On remarque que chaque cluster à un pourcentage d'inertie expliquée assez élevé.

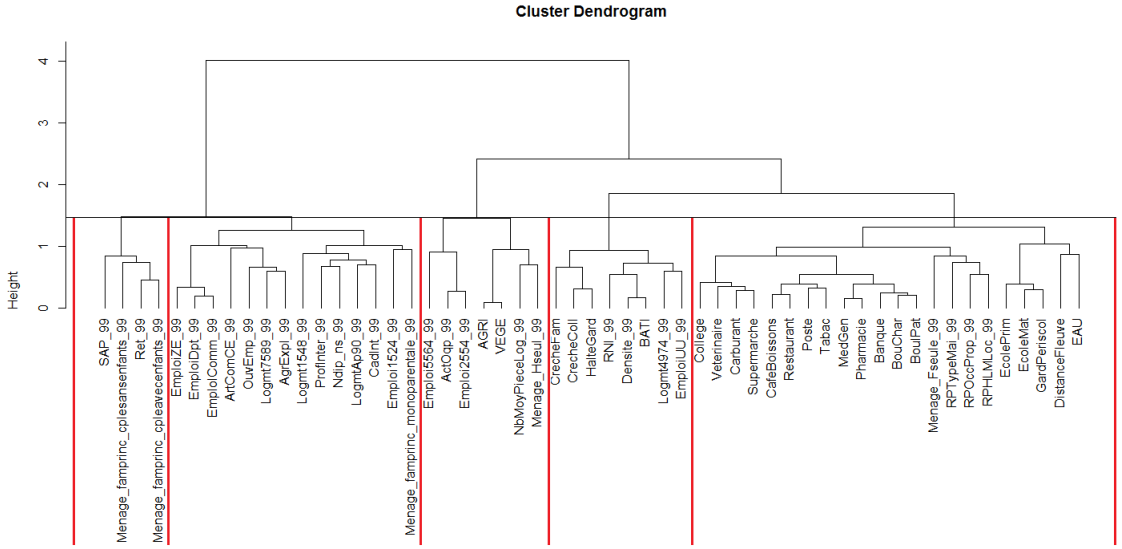



FIGURE 3.3 – Dendrogramme de hclustvar sur les données de 1999.

Les résultats rassemblés dans la Table 3.4 détaillent les variables contenues dans chaque cluster. Pour chaque variable, son "squared loading" avec la variable synthétique du cluster est donné. Pour les variables quantitatives, la corrélation avec la variable synthétique est également indiquée. Ce tableau permet d'interpréter les variables synthétiques et de leur donner ainsi un sens en tant qu'indicateurs composites de qualité de vie.

Tout d'abord, on remarque que la proportion de territoires agricoles (mesurée à l'échelle municipale) est fortement corrélée avec le nombre moyen de pièces par logement et le taux d'actifs occupés. Cela forme le premier cluster de variables (Cluster 1).

Cela montre que jusqu'à 1999, les actifs vivaient principalement dans des communes où l'activité principale était l'agriculture. Nous appelons ce premier cluster (et donc l'indicateur de qualité de vie associé) "Environnement Naturel et Taux d'Emploi".

Le second cluster de variables (Cluster 2) décrit les conditions d'emplois des résidents (lieu, type d'emplois) et l'ancienneté des logements au sein de la commune. La variable synthétique associée à ce cluster est corrélée négativement avec la proportion de résidents travaillant à l'extérieur de leur commune de résidence. Elle est aussi corrélée négativement avec le nombre de logements récents (construits entre les années 1970 et 1990) de la commune. De plus, cette variable synthétique est corrélée positivement avec la proportion d'agriculteurs (parmi tous les emplois au sein de la commune) et la proportion d'emplois au sein de la commune. Cette seconde variable synthétique est appelée "Accès à l'Emploi".

Le troisième groupe de variables (Cluster 3) reflète le processus d'urbanisation. Logiquement ce groupe associe le pourcentage de batiments sur la commune avec la densité de population. On remarque également que ces variables sont liées à la présence de crèches et de garderies sur la commune. Ce cluster est également caractérisé par la proportion de résidents travaillant dans la même unité urbaine que leur commune de résidence, mais également à la variable concernant le revenu moyen des habitants. L'indicateur de qualité de vie associée est appelé "Environnement Socio-Economique Urbain".

Le quatrième cluster de variables (Cluster 4) montre une corrélation positive entre la proportion de retraités et la proportion de ménages sans enfants sur la commune. On observe également une corrélation négative entre la variable synthétique de ce cluster et le nombre de personnes sans emplois ainsi que le nombre de ménages avec des enfants. La variable synthétique associée à ce cluster est appelée "Structure des Ménages et Modes de Vie".

Le dernier cluster de variables (Cluster 5) rassemble toutes les variables qualitatives liées à la présence et à l'accès aux services au sein de la commune et ceci sans distinction entre les services publics et les services commerciaux. La variable synthétique associée est appelée "Disponibilité et Accès aux Services".

Pour donner plus de sens à ces variables synthétiques, on les regarde comme des gradients. Ainsi pour chaque indicateur composite on récapitule quelles variables sont le plus associées à une valeur négative ou positive de l'indicateur. Cette interprétation est donnée à la Table 3.5. On peut également représenter les valeurs d'un indicateur composite sur notre zone d'étude, en le découpant en classes selon ses quantiles par exemple. La Figure 3.4 représente l'indicateur composite "Environnement Naturel et Taux d'Emploi" sur la zone d'étude. On voit par exemple que la zone la plus claire au sud de la Garonne entre Bordeaux et Agen correspond à des communes ayant une forte

| Cluster 1 : Environnement Naturel et Taux d'Emploi | | |
|---|---|---|
| *Variables* | *Squared loading* | *Corrélations (pour variables quanti)* |
| Agri | 0.62 | 0.79 |
| Vege | 0.57 | -0.75 |
| NbMoyPieceLog_99 | 0.46 | 0.68 |
| ActOqp_99 | 0.39 | 0.62 |
| Emploi2554_99 | 0.35 | 0.59 |
| Menage_Hseul_99 | 0.16 | -0.40 |
| Emploi5564_99 | 0.11 | 0.34 |
| **Cluster 2 : Accès à l'Emploi** | | |
| *Variables* | *Squared loading* | *Corrélations (pour variables quanti)* |
| EmploIComm_99 | 0.69 | 0.83 |
| EmploiDpt_99 | 0.68 | -0.82 |
| EmploiZE_99 | 0.60 | -0.78 |
| Logmt7589_99 | 0.43 | -0.66 |
| AgrExpl_99 | 0.36 | 0.60 |
| LogmtAp90_99 | 0.33 | -0.57 |
| ProfInter_99 | 0.30 | -0.55 |
| Ndip_ns_99 | 0.21 | 0.45 |
| CadInt_99 | 0.18 | -0.42 |
| OuvEmp_99 | 0.12 | -0.35 |
| Logmt1548_99 | 0.07 | 0.26 |
| Emploi1524_99 | 0.03 | 0.17 |
| ArtComCE_99 | 0.01 | -0.08 |
| Menage_famprinc_monoparentale_99 | 0.00 | 0.02 |
| **Cluster 3 : Environnement Socio-Economique Urbain** | | |
| *Variables* | *Squared loading* | *Corrélations (pour variables quanti)* |
| BATI | 0.78 | 0.88 |
| Densite_99 | 0.64 | 0.80 |
| CrecheColl | 0.55 | - |
| HalteGard | 0.51 | - |
| EmploiUU_99 | 0.44 | 0.66 |
| RNI_99 | 0.41 | 0.65 |
| Logmt4974_99 | 0.38 | 0.62 |
| CrecheFam | 0.35 | - |
| **Cluster 4 : Structure des Ménages et Modes de Vie** | | |
| *Variables* | *Squared loading* | *Corrélations (pour variables quanti)* |
| Ret_99 | 0.66 | 0.81 |
| Menage_famprinc_cpleavecenfants_99 | 0.65 | -0.81 |
| Menage_famprinc_cplesansenfants_99 | 0.38 | 0.61 |
| SAP_99 | 0.29 | -0.54 |
| **Cluster 5 : Disponibilité et Accessibilité des services** | | |
| *Variables* | *Squared loading* | *Corrélations (pour variables quanti)* |
| BoulPat | 0.78 | - |
| Pharmacie | 0.76 | - |
| BouChar | 0.74 | - |
| Tabac | 0.70 | - |
| MedGen | 0.69 | - |
| Carburant | 0.69 | - |
| CafeBoissons | 0.67 | - |
| Banque | 0.67 | - |
| Restaurant | 0.63 | - |
| Poste | 0.59 | - |
| Veterinaire | 0.51 | - |
| College | 0.50 | - |
| Supermarche | 0.49 | - |
| EcolePrim | 0.45 | - |
| EcoleMat | 0.41 | - |
| RPHLMLoc_99 | 0.36 | - |
| GardPeriscol | 0.35 | - |
| RPOccProp_99 | 0.28 | -0.53 |
| RPTypeMai_99 | 0.21 | - |
| Menage_Fseule_99 | 0.07 | 0.27 |
| EAU | 0.05 | 0.23 |
| DistanceFleuve | 0.02 | -0.13 |

TABLE 3.4 – Liaison entre les variables d'origine et la variable synthétique pour chaque cluster de variables de l'année 1999.

proportion de territoires végétalisés, ces communes appartiennent majoritairement à la forêt des Landes. Ceci est également le cas au sud de la zone où l'on retrouve des communes montagnardes des Pyrénées.

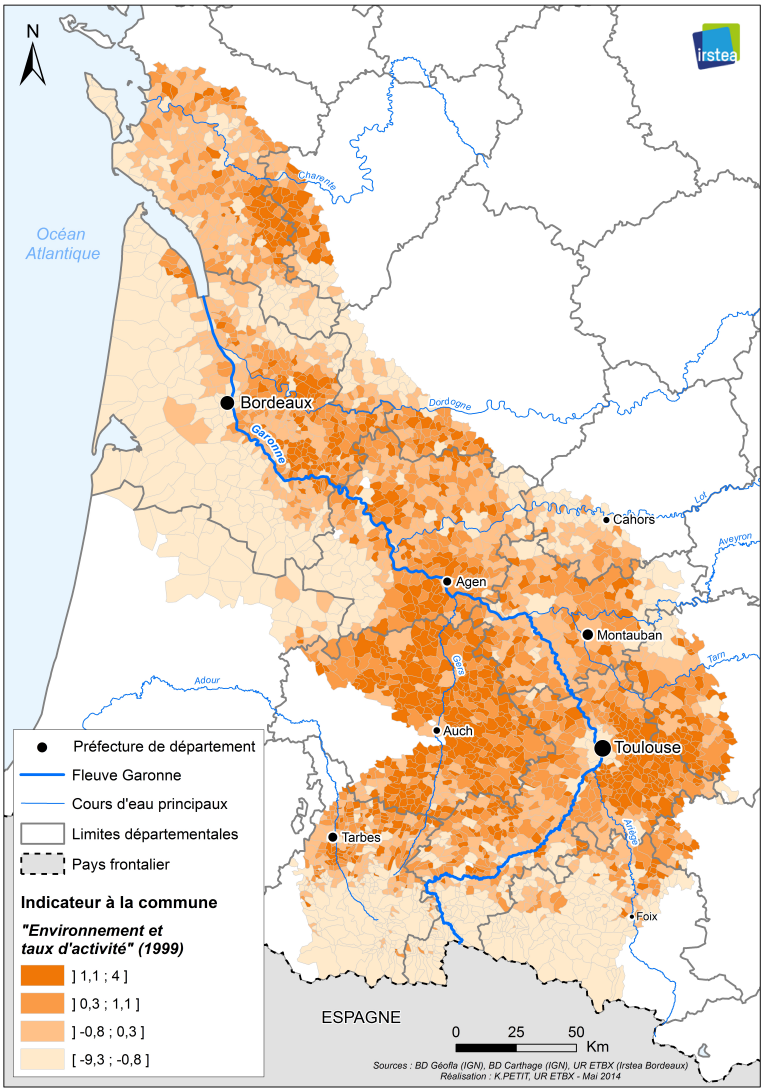



FIGURE 3.4 – Représentation de l'indicateur composite "Environnement Naturel et Taux d'Emploi" en 1999 découpé par la méthode des quantiles.

Par la suite, nous allons créer une typologie des communes afin de rassembler les communes qui prennent les mêmes valeurs sur les différents indicateurs composites construits.

#### 3.3.2.2 Typologie des communes sur les indicateurs composites de 1999

Nous utilisons ici les 5 indicateurs composites précédemment construits afin d'établir une typologie des communes. Nous utilisons pour cela une classification ascendante

| **Indicateur composite** | **Valeurs négatives** | **Valeurs positives** |
|---|---|---|
| Environnement Naturel et Taux d'Emploi | Forte proportion de territoires forestiers<br>Faible taux d'emploi | Forte proportion de terres agricoles<br>Taux d'emploi élevé |
| Accès à l'Emploi | Emploi au sein du département<br>Faible proportion d'agriculteurs<br>Logements construits après les années 1970 | Emploi dans la commune de résidence<br>Proportion importante d'agriculteurs |
| Environnement Socio-Economique Urbain | Faible proportion de batiments<br>Faible densité de population<br>Faibles revenus | Forte proportion de batiments<br>Forte densité de population<br>Revenus élevés<br>Emploi au sein de l'unité urbaine de résidence<br>Services de petite enfance<br>Logements construits entre 1950 et 1975 |
| Structure des Ménages et Modes de Vie | Faible proportion de retraités<br>Proportion importante de couples avec enfants<br>Faible proportion de personnes en activité | Proportion importante de retraités<br>Proportion importante de couples sans enfants<br>Proportion importante de personnes en activité |
| Disponibilité et Accessibilité des Services | Peu de services sur la commune<br>Logements occupés par les propriétaires | Nombre important de services<br>Logements locatifs |

Table 3.5 – Lecture des indicateurs composites de QLV en 1999.

hiérarchique avec critère de Ward sur les valeurs prises par les communes sur les indicateurs. Cette classification d'observations a pour but de rassembler dans une même classe (on rappelle que le mot cluster est réservé aux clusters de variables) les communes qui se ressemblent. Nous avons retenu une partition en 5 classes de communes, ce choix du nombre de classe a été effectué afin d'avoir des profils de communes distincts et interprétables. L'interprétation des classes créées est réalisée à l'aide de la Table 3.6 qui donne les valeurs moyennes des différents indicateurs au sein de chaque classe.

| **Indicateur composite** | **Classe de communes** | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| | *n=476* | *n=1188* | *n=1010* | *n=458* | *n=155* |
| Environnement Naturel et Taux d'Emploi | **-2.53** | **0.84** | **0.56** | -0.56 | -0.66 |
| Accès à l'Emploi | 0.4 | **1.45** | **-1.86** | 0.49 | **-1.68** |
| Environnement Socio-Economique Urbain | -0.67 | **-0.79** | -0.15 | 0.63 | **7.22** |
| Structure des Ménages et Modes de Vie | **0.77** | **0.44** | **-0.83** | 0.14 | -0.71 |
| Disponibilité et Accessibilité des Services | **-1.41** | **-1.75** | **-0.92** | **5.36** | **7.91** |

En gras : valeurs significativement différentes de la moyenne globale de l'indicateur (par construction la moyenne globale est nulle) ; p-value inférieure à $10^{-3}$.

TABLE 3.6 – Moyenne des indicateurs composites sur les cinq classes de communes créées en 1999.

Le parallèle entre la Table 3.5 et la Table 3.6 permet d'interpréter simplement les caractéristiques des classes de communes. Pour cela, on compare la valeur moyenne des indicateurs au sein des différentes classes à la valeur moyenne de l'indicateur sur l'ensemble des communes (égale à 0 par construction). La Table 3.6 montre en gras les valeurs significativement différentes de 0. Ce test n'a pas une grande signification statistique car les indicateurs composites ont été utilisés pour créer les classes de communes. Cependant, il est utile afin d'identifier les indicateurs caractérisant chaque classe de communes. Ceci amène à la description des classes de communes suivantes :

- **La classe 1** contient 476 communes où le mode de vie ressemble à celui des retraités. En effet, l'indicateur composite "Structure des Menages et Modes de Vie" pour cette classe de communes possède une valeur moyenne positive, cela implique une proportion importante de retraités. On observe également que pour les communes de cette classe, l'emploi est principalement situé au sein de la commune, cela est observé grâce à la valeur moyenne positive de l'indicateur composite "Accès à l'Emploi". Au vu de la valeur moyenne négative de l'indicateur "Disponibilité et Accessibilité des Services", les communes de la classe 1 possèdent peu de services. Ces communes sont plus adaptées pour des activités forestières car la proportion de forêts sur ces communes est supérieure à la moyenne. On observe également un taux d'emploi inférieur à la moyenne (valeur moyenne négative de l'indicateur "Environnement Naturel et Taux d'Emploi").

- **La classe 2** contient 1188 communes dont la proportion de terres agricoles est supérieure à la moyenne. Cette proportion de terres agricoles est associée à un nombre important d'exploitants agricoles travaillant dans leur commune de résidence. Ces communes sont également un lieu de résidence pour les retraités et les couples sans enfants. Comme pour les communes de la classe 1, on remarque un faible accès aux services.
- **La classe 3** rassemble 1010 communes principalement caractérisées par leur accès à l'emploi. En effet, une faible valeur moyenne de l'indicateur "Accès à l'Emploi" sur cette classe indique que les habitants de ces communes travaillent principalement hors de leur commune de résidence. Les ménages sont principalement composés de couples avec enfants vivant dans des logements assez récents. Ces communes ont une offre de services limitée.
- **La classe 4** contient 458 communes où les conditions de vie sont principalement guidées par un accès aux services important. Ces communes ont également une proportion de bâtiments importante et une densité de population assez élevée. Le revenu des habitants est supérieur à la moyenne. Ces communes sont principalement situées à la périphérie des deux pôles urbains de notre zone d'étude que sont Bordeaux et Toulouse.
- **La classe 5** regroupe 155 communes urbaines et péri-urbaines avec une très forte densité de population et une forte proportion de bâtiments. Les revenus de la population sont nettement supérieurs à la moyenne et les habitants travaillent principalement hors de leur commune de résidence. On remarque également une offre de services supérieure à celle de toutes les autres classes et une forte proportion de logement construits entre les années 1950 et 1975. Ces communes correspondent aux deux zones urbaines de Bordeaux et Toulouse.

### 3.3.3 Résultats de hclustvar et typologie des communes pour l'année 2009

#### 3.3.3.1 Construction des indicateurs composites de l'année 2009 avec hclustvar

L'étude établie sur les données de 1999 a été reproduite sur les données de 2009 sur la même zone d'étude. Suite à l'application de la méthode hclustvar, nous avons choisi de retenir 5 clusters de variables. Ce choix a été effectué en fonction du dendrogramme mais également afin de comparer les indicateurs composites qui en découlent à ceux obtenus pour l'année 1999. Les "squared loadings" entre les variables de chaque cluster et la variable synthétique associée sont donnés dans la Table 3.7. La Table 3.8 contient l'aide à la lecture des variables synthétiques (indicateurs composites) en fonction de

leurs valeurs (positives ou négatives) sur les différentes communes. L'examen des liaisons entre les variables initiales de chaque cluster avec la variable synthétique associée révèle différents changements.

Premièrement, une forte proportion de territoires agricoles est maintenant corrélée avec une forte proportion de résidences principales. Ces deux variables sont corrélées positivement à la variable synthétique du cluster 1, ce cluster est donc nommé : "Environnement Naturel et Résidences Principales". On remarque que contrairement à l'année 1999, l'activité agricole n'est plus liée au taux d'emploi et que les communes agricoles sont dorénavant plus tournées vers le développement de logements. Cela peut être expliqué par l'attractivité paysagère de ces territoires agricoles.

Le second cluster de variables rassemble les variables caractérisant le taux d'emplois des résidents et la proportion des 25-54 ans ayant un emploi. La variable synthétique associée est corrélée positivement à ces deux variables et elle peut être nommée "Taux d'Emploi".

La variable synthétique associée au troisième cluster de variables est fortement corrélée avec la proportion d'habitants ayant un emploi au sein de leur commune de résidence mais également avec la proportion d'agriculteurs sur la commune. Inversement, la variable synthétique est corrélée négativement à la fois avec la proportion d'habitants travaillant au sein de leur département de résidence et avec la proportion d'ouvriers et employés. Cette troisième variable synthétique est appelée "Accès à l'Emploi".

La variable synthétique du quatrième cluster est corrélée négativement à la proportion de retraités sur la commune. Inversement, elle est fortement corrélée positivement avec la proportion de couples avec enfants mais également avec le revenu et avec la proportion d'emplois intermédiaires et qualifiés. Cet indicateur composite (variable synthétique) est appelé "Modes de Vie et Niveaux de Vie".

Le cinquième et dernier cluster de variables caractérise les conditions urbaines des communes : la proportion de zones bâties, la présence de services, la densité de population et différentes variables caractérisant les types de logements. L'indicateur composite associé est appelé "Services et Conditions Urbaines".

#### 3.3.3.2 Typologie des communes sur les indicateurs composites de 2009

Comme pour l'année 1999, nous créons ici une typologie des communes en se basant sur leurs valeurs sur les 5 indicateurs composites construits à l'aide de hclustvar pour l'année 2009. Nous avons retenu une partition en 5 classes de communes afin de pouvoir la comparer à la typologie de 1999 et ainsi établir des trajectoires d'évolution entre les profils de communes. Afin d'interpréter les classes créées nous avons reporté dans la Table 3.9 les valeurs moyennes des indicateurs au sein de chaque classe.

Ceci amène à la description suivante :

| Cluster 1 : Environnement Naturel et Résidences Principales | | |
|---|---|---|
| *Variables* | *Squared Loadings* | *Corrélations (pour variables quanti)* |
| VEGE | 0.90 | -0.95 |
| AGRI | 0.85 | 0.92 |
| RPLogTot_09 | 0.57 | 0.76 |
| ArtComCE_09 | 0.03 | -0.17 |
| **Cluster 2 : Taux d'Emploi** | | |
| *Variables* | *Squared Loadings* | *Corrélations (pour variables quanti)* |
| Emploi2554_09 | 0.83 | 0.91 |
| ActOqp_09 | 0.75 | 0.87 |
| SAP_09 | 0.16 | -0.39 |
| Emploi1524_09 | 0.04 | 0.20 |
| **Cluster 3 : Accès à l'Emploi** | | |
| *Variables* | *Squared Loadings* | *Corrélations (pour variables quanti)* |
| EmploiComm_09 | 0.71 | 0.84 |
| EmploiDpt_09 | 0.68 | -0.83 |
| AgrExpl_09 | 0.39 | 0.62 |
| OuvEmp_09 | 0.26 | -0.51 |
| **Cluster 4 : Modes de Vie et Niveaux de Vie** | | |
| *Variables* | *Squared Loadings* | *Corrélations (pour variables quanti)* |
| Ret_09 | 0.57 | -0.76 |
| Menage_famprinc_cpleavecenfants_09 | 0.54 | 0.74 |
| RNIMoy_09 | 0.44 | 0.70 |
| CadInt_09 | 0.41 | 0.64 |
| ProfInter_09 | 0.37 | 0.61 |
| Ndip_ns_09 | 0.20 | -0.45 |
| Emploi5564_09 | 0.17 | 0.41 |
| Menage_famprinc_cplesansenfants_09 | 0.15 | -0.39 |
| Menage_Fseule_09 | 0.13 | -0.36 |
| Menage_Hseul_09 | 0.03 | -0.18 |
| Menage_famprinc_monoparentale_09 | 0.00 | 0.03 |
| **Cluster 5 : Services et Conditions Urbaines** | | |
| *Variables* | *Squared Loadings* | *Corrélations (pour variables quanti)* |
| Pharmacie | 0.86 | - |
| MedOmni | 0.81 | - |
| ChiDentiste | 0.80 | - |
| Boul | 0.75 | - |
| BanqueCE | 0.72 | - |
| BouChar | 0.69 | - |
| EcolElem | 0.68 | - |
| Restaurant | 0.66 | - |
| EcoleMat | 0.66 | - |
| Supermarche | 0.63 | - |
| College | 0.61 | - |
| Veterinaire | 0.60 | - |
| Poste | 0.57 | - |
| BATI | 0.47 | 0.69 |
| GardPrescol | 0.43 | - |
| RPTypeMai_09 | 0.42 | - |
| RPOccProp_09 | 0.41 | -0.64 |
| RPHLMLoc_09 | 0.38 | - |
| Densite_09 | 0.32 | 0.56 |
| Superette | 0.29 | - |
| Epicerie | 0.24 | - |
| EAU | 0.04 | 0.19 |
| DistanceFleuve | 0.02 | -0.15 |

TABLE 3.7 – Liaison entre les variables d'origine et la variable synthétique pour chaque cluster de variables de l'année 2009.

| **Indicateur composite** | **Valeurs négatives** | **Valeurs positives** |
|---|---|---|
| Environnement Naturel et Résidences Principales | Forte proportion de territoires forestiers ou végétaux<br>Faible proportion de résidences principales | Forte proportion de territoires agricoles<br>Forte proportion de résidences principales |
| Taux d'Emploi | Faible taux d'emploi | Taux d'emploi élevé |
| Accès à l'Emploi | Emploi au sein du département<br>Faible proportion d'agriculteurs<br>Proportion importante d'emplois intermédiaires | Emploi dans la commune de résidence<br>Forte proportion d'agriculteurs<br>Faible proportion d'emplois intermédiaires |
| Modes de Vie et Niveaux de Vie | Proportion importante de retraités<br>Faible proportion de couples avec enfants<br>Faible revenu<br>Faible proportion d'emplois très qualifiés<br>Faible proportion d'ouvriers | Faible proportion de retraités<br>Proportion élevée de couples avec enfants<br>Revenu important<br>Proportion élevée d'emplois qualifiés<br>Proportion élevée d'ouvriers |
| Services et Conditions Urbaines | Peu de services<br>Faible proportion de bâti<br>Faible densité de population<br>Forte proportion de logements occupés par leur propriétaire | Offre importante de services<br>Proportion importante de bâti<br>Forte densité de population<br>Faible proportion de logements occupés par leur propriétaire |

TABLE 3.8 – Lecture des indicateurs composites de QLV en 2009.

| **Indicateur composite** | **Classe de communes** | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| | *n=460* | *n=1095* | *n=1162* | *n=318* | *n=252* |
| Environnement Naturel et Résidences Principales | **-2.96** | **0.44** | **0.69** | 0.28 | -0.06 |
| Taux d'Emploi | **-0.67** | -0.01 | **0.52** | -0.30 | -0.74 |
| Accès à l'Emploi | 0.01 | **0.90** | **-0.90** | -0.23 | 0.50 |
| Modes de Vie et Niveaux de Vie | **-0.76** | **-1.08** | **1.37** | -0.08 | -0.12 |
| Services et Conditions Urbaines | -1.2 | **-1.57** | **-1.31** | **3.95** | **10.08** |

En gras : valeurs significativement différentes de la moyenne globale de l'indicateur (par construction la moyenne globale est nulle) ; p-value inférieure à $10^{-3}$.

TABLE 3.9 – Moyenne des indicateurs composites pour les 5 classes de communes de l'année 2009.

- **La classe 1** contient 460 communes principalement caractérisées par leur forte proportion de territoires forestiers. Ces communes ont un taux d'emplois assez faible ainsi qu'une proportion de résidences principales inférieure à la moyenne. La population de retraités est assez importante. Les revenus des habitants sont plutôt faibles et ils possèdent peu d'emplois qualifiés.
- **La classe 2** rassemble 1095 communes caractérisées par la présence de territoires agricoles. Ces communes ont une proportion importante de retraités et d'emplois agricoles. Les habitants travaillent principalement au sein de leur commune de résidence. Ces communes offrent très peu de services à leurs habitants.
- **La classe 3** contient 1162 communes contenant une proportion importante de territoires agricoles. Malgré le très faible accès aux services, la proportion de résidences principales est importante et on y trouve beaucoup de couples avec enfants. Les habitants ont des revenus assez importants, des emplois relativement qualifiés et travaillent majoritairement hors de leur commune de résidence (au sein du département).
- **La classe 4** rassemblant 318 communes est essentiellement caractérisée par une forte présence de services et une proportion élevée de logements locatifs. Ces communes sont principalement situées autour des pôles urbains.
- **La classe 5** contient 252 communes urbaines. Ces communes ont une offre de services extrêmement importante ainsi qu'une densité de population et une proportion de bâtiments très élevée (plus élevée que la classe 4). De plus, la proportion d'habitants travaillant dans leur commune de résidence est plus importante que dans la classe 4.

Les typologies des communes de 1999 et 2009 sont représentées sur les cartes géographiques de la zone d'étude de la Figure 3.5.

### 3.3.4 Trajectoires des communes entre 1999 et 2009

Les résultats présentés dans cette section sont déduits de l'interprétation de la Table 3.10 et des cartes présentées à la Figure 3.5 qui montrent l'évolution des typologies de communes entre 1999 et 2009. De plus les deux cartes présentées à la Figure 3.6 montrent quelles sont les communes ayant changé de classe entre 1999 et 2009 (passage de la classe 2 à la classe 3 et passage de la classe 4 à la classe 5).

Une des tendances majeures dans le processus de développement des communes étudiées est l'évolution de l'usage des territoires forestiers et agricoles. L'analyse des différences entre les typologies de 1999 et 2009 montrent que ces variables, qui reflètent les conditions environnementales des communes, sont corrélées avec les dynamiques socio-économiques à l'échelle des communes. En 1999, les territoires agricoles étaient principalement des zones d'emplois et de production. En 2009, ces communes

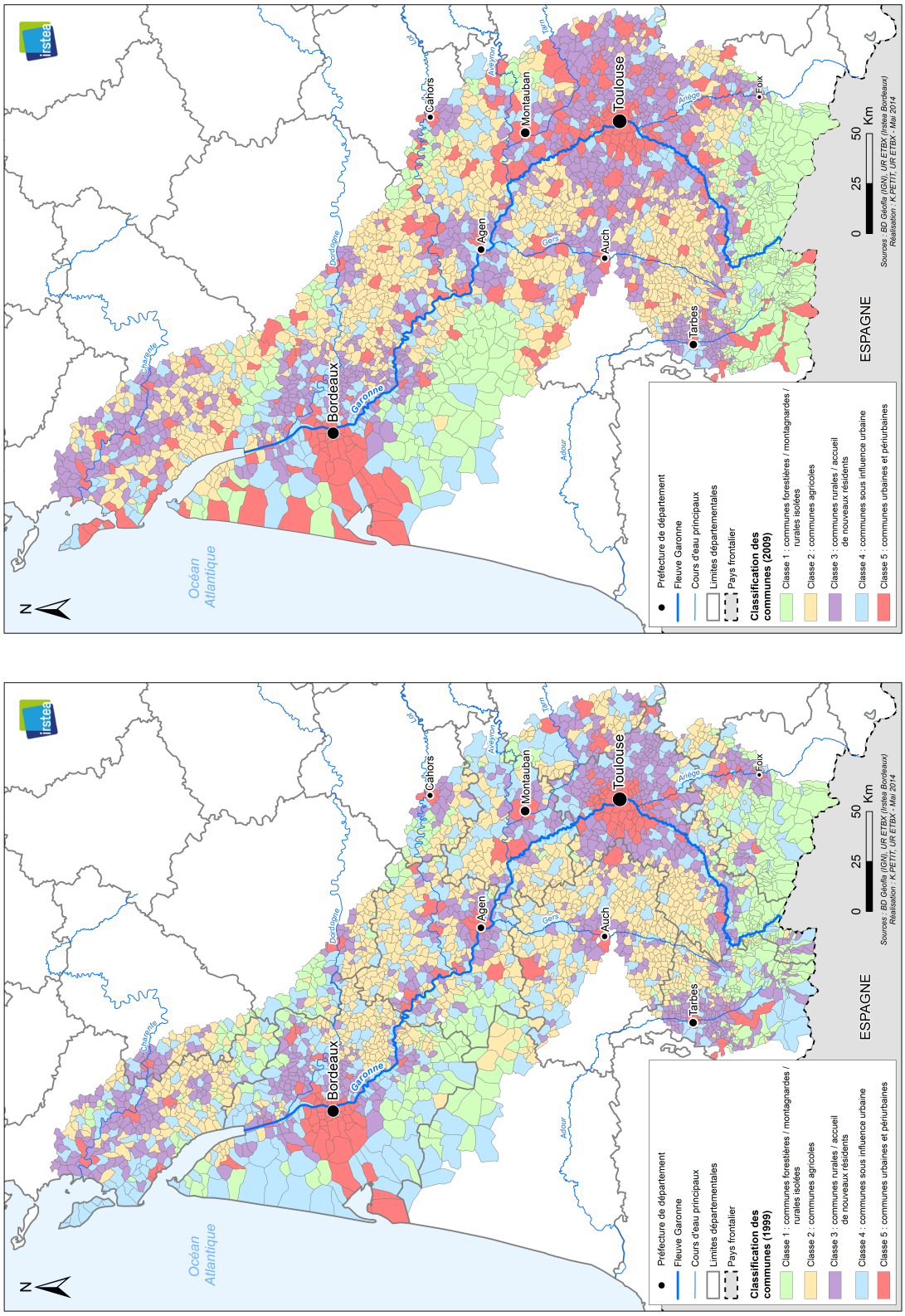



FIGURE 3.5 – Carte des typologies des communes de 1999 (en bas) et de 2009 (en haut).

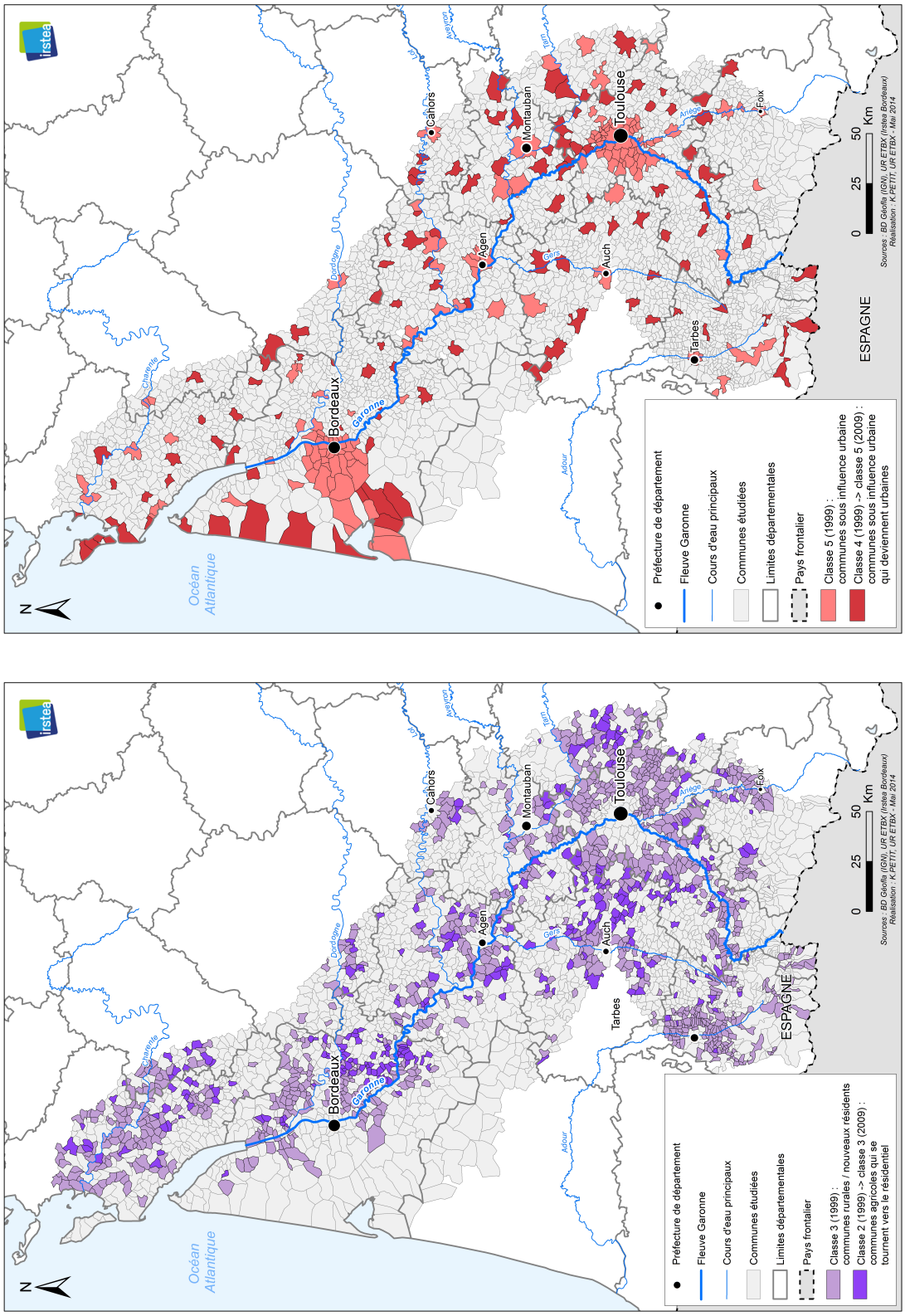



FIGURE 3.6 – Communes passant de la classe 2 en 1999 à la classe 3 en 2009 (en bas). Communes passant de la classe 4 en 1999 à la classe 5 en 2009 (en haut).

| | | Classe en 2009 | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| | 1 | 344 | 105 | 27 | 0 | 0 |
| **Classe** | 2 | 28 | 845 | 315 | 0 | 0 |
| **en** | 3 | 56 | 108 | 801 | 44 | 1 |
| **1999** | 4 | 32 | 37 | 18 | 252 | 119 |
| | 5 | 0 | 0 | 1 | 22 | 132 |

TABLE 3.10 – Tableau croisé des classes de communes de 1999 et 2009.

deviennent plus attractives pour les aménités paysagères qu'elles procurent aux nouveaux résidents, cela est encore plus marqué dans le cas des communes peu éloignées des pôles urbains. Bien que ce phénomène ne soit pas spécifique à la zone étudiée, nous pouvons remarquer certaines spécificités. Les communes situées dans des zones montagneuses, forestières, ou dans des zones agricoles proches des berges de l'estuaire de la Gironde ne sont pas affectées par le phénomène d'urbanisation. Le manque de services sur ces communes couplé à la forte proportion de retraités implique une plus grande vulnérabilité face aux risques associés aux changements climatiques (tempêtes, inondations). Ces communes appartenaient aux classes 1 et 2 en 1999 et restent majoritairement dans les classes 1 et 2 de 2009.
Les communes agricoles appartenant à la classe 2 en 1999 se retrouvant dans la classe 3 en 2009 sont attractives grâce à leur proximité à des villes de taille moyenne proposant des emplois dans des catégories intermédiaires. En 1999, les communes de la classe 3 étaient assez proches du fleuve. En 2009, on assiste à l'étalement des petites villes et on retrouve des communes de la classe 3, ayant le même profil, dans des zones plus éloignées du fleuve. Initialement ces communes étaient habitées par des populations moins qualifiées et moins diplômées. Le développement résidentiel entre 1999 et 2009 participe à l'augmentation de la population au sein des banlieues de Toulouse et Bordeaux.
On observe également qu'un tiers des communes de la classe 4 en 1999, se retrouvent dans la classe 5 en 2009. Cela est par exemple le cas des communes situées entre Bordeaux et la côte Atlantique mais aussi des communes situées entre Toulouse et la ville moyenne de Montauban. Cette transition est en partie dûe à la plus grande disponibilité des emplois et à une augmentation des investissements concernant les services de proximité.

## 3.4 Conclusion

Ce chapitre a permis de présenter la méthode de classification de variables hclustvar. Cette approche permet de construire des clusters de variables sans a priori sur

leurs relations. De plus l'obtention de variables synthétiques liées à chaque cluster de variables permet de réduire considérablement la dimension des données. La méthode a été testée sur des communes de la ZGG sur deux années différentes. Les résultats obtenus ont permis de comprendre la structuration des dimensions de la qualité de vie sur ces deux années. Par la suite une classification des communes combinée à une représentation cartographique a confirmé l'hétérogénéité des conditions de vie sur la ZGG. La comparaison de deux années avec un intervalle de temps de 10 ans à permis d'observer les différences de structuration des composantes de la qualité de vie mais également de mieux comprendre l'évolution des profils de communes. Les principales trajectoires d'évolutions des communes ont permis d'aboutir aux conclusions suivantes. On peut observer d'une part que le phénomène de métropolisation peut amplifier à moyen terme la vulnérabilité socio-économique de certaines communes et que la densification des communes, sous l'influence des deux pôles urbains autour de Toulouse et de Bordeaux, a permis une amélioration de l'offre de services et d'emplois dans ces communes. D'autre part, les communes rurales localisées loin des centres d'emplois et proposant une très faible offre de services restent isolées. Si de nouveaux résidents ne viennent pas dans ces communes, leur population deviendra plus vieillissante et plus vulnérable aux risques dûs aux changements climatiques. Une option envisageable pour contrer les risques liés au processus d'urbanisation des grandes villes est de promouvoir la polycentricité des ensembles urbains. Le développement des villes de taille moyenne par la création d'emplois et d'accès aux services joue un rôle central. Les communes rurales situées proches de ces villes de taille moyennes pourront ainsi se développer plus aisément, cependant cela se fera probablement aux dépends des territoires agricoles, ce qui peut représenter une menace pour les écosystèmes associés à ce type de territoire.

# 4 Analyse factorielle multiple de données mixtes : la méthode MFAmix

## Sommaire

## 4.1 Introduction

L'analyse factorielle multiple (AFM) est une méthode d'analyse factorielle qui permet de prendre en compte le fait que les observations sont décrites par des variables naturellement structurées en groupes ou thématiques. Initialement l'AFM a été mise en place pour l'analyse de groupes de variables quantitatives, voir par exemple Escofier and Pagès (1983). Elle a ensuite été élargie à l'analyse de groupes de variables qualitatives, voir Pagès (2002) puis à l'étude d'un tableau de données que l'on qualifiera de "semi-mixte", où chaque groupe peut être soit de type quantitatif, soit de type qualitatif, voir Bécue-Bertaut and Pagès (2008). Cette méthode se distingue d'une analyse en composantes principales (ACP) ou d'une analyse des correspondances multiples (ACM) globale appliquée à l'ensemble des données dans la mesure où elle permet de prendre en compte la structure en groupes de l'ensemble des variables. Pour cela, l'AFM applique une pondération particulière aux variables selon leur appartenance aux différents groupes. Cette pondération consiste à diviser chaque variable par la racine de la valeur propre issue de l'ACP du groupe de variables quantitatives (resp. de l'ACM du groupe de variables qualitatives) auquel elle appartient. Ainsi l'influence des groupes est équilibrée dans la construction des composantes principales globales, contrairement à une méthode d'analyse factorielle classique qui ne considère pas la structure en groupes des variables et qui accorderait plus d'importance à un groupe avec une structure forte ou à un groupe de grande dimension. Ainsi, avec les méthodes classiques d'analyse factorielle, l'obtention des composantes principales serait influencée de manière prépondérante par ce type de groupe de variables et celles-ci ne résumeraient pas de manière objective l'information apportée par l'ensemble des données. De plus, l'AFM permet de situer les différents groupes dans un même référentiel, en vue de leur comparaison, ce qui n'est pas permis par des analyses factorielles classiques qui seraient réalisées de manière indépendante sur chaque groupe. L'écriture actuelle de l'AFM et son implémentation dans le package R `FactoMineR`, voir Husson et al. (2015), ne permettant pas d'intégrer des groupes de variables mixtes dans l'analyse, nous proposons une extension de l'AFM, appelée MFAmix, qui permet l'analyse de groupes de variables mixtes (groupes comportant des variables quantitatives et des variables qualitatives). Nous proposons également une écriture de la méthode sous forme de GSVD, permettant notamment d'exprimer les coordonnées factorielles sous forme matricielle. Nous allons voir dans ce chapitre comment mettre en œuvre la méthode MFAmix et quelles sont les principales sorties numériques disponibles au sein de cette méthode. Le principe général repose essentiellement sur deux étapes. Tout d'abord, on analyse chaque groupe pris séparément avec la méthode PCAmix. On obtient ainsi la plus grande valeur propre associée à chaque groupe de variables. Puis, on applique PCAmix sur l'ensemble des variables (union des différents groupes) où chaque variable

est pondérée par l'inverse de la racine de la première valeur propre du groupe dont elle est issue. Ainsi l'influence de chaque groupe est équilibrée dans la construction des composantes principales globales.
La Section 4.2 présente la méthode MFAmix ainsi que les sorties numériques associées. Un exemple illustratif de la méthode, à l'aide du package `PCAmixdata`, sera effectué sur des données socio-économiques relatives à la qualité de vie d'un ensemble de communes de la Gironde. La Section 4.3 décrit une méthode permettant de construire des nouvelles composantes principales avec MFAmix, ces nouvelles composantes principales seront calculées sur un nombre restreint de variables, tout en étant fortement corrélées aux composantes principales calculées sur l'ensemble des variables. Finalement, la Section 4.4 montre comment la méthode MFAmix peut être utilisée pour la construction d'indicateurs composites de qualité de vie.

## 4.2 La méthode MFAmix

Dans le cas de MFAmix, le tableau de données $\mathbf{X}$ que l'on veut analyser comporte $n$ observations décrites par $p$ variables, les $p$ variables sont séparées en $G$ groupes. Au sein d'un groupe $g$ les variables peuvent êtres quantitatives et/ou qualitatives. Chaque groupe $g$ est représenté par la matrice de données $\mathbf{X}^{(g)} = [\mathbf{X}_1^{(g)}, \mathbf{X}_2^{(g)}]$ où $\mathbf{X}_1^{(g)}$ (resp. $\mathbf{X}_2^{(g)}$) est la sous-matrice de $\mathbf{X}^{(g)}$ contenant les $p_1^{(g)}$ variables quantitatives (resp. les $p_2^{(g)}$ variables qualitatives). On peut alors réécrire la matrice $\mathbf{X}$ de la manière suivante : $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ où $\mathbf{X}_1 = [\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_1^{(G)}]$ et $\mathbf{X}_2 = [\mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_2^{(G)}]$. On note $\mathbf{Z}$, $\mathbf{N}$ et $\mathbf{M}$ les matrices construites avec $\mathbf{X}_1$ et $\mathbf{X}_2$ comme décrit à la Section 2.3.1. Ainsi, on obtient $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{Z}_2]$ où $\mathbf{Z}_1$ est la matrice $\mathbf{X}_1$ centrée-réduite et $\mathbf{Z}_2$ est la matrice centrée des indicatrices des $m$ modalités de $\mathbf{X}_2$, $\mathbf{N} = \frac{1}{n}\mathbb{I}_n$ et $\mathbf{M}$ est la matrice diagonale des poids des colonnes de $\mathbf{Z}$. Les $p_1 = \sum_{g=1}^{G} p_1^{(g)}$ premières colonnes sont pondérées par 1 et les $m$ dernières colonnes sont pondérées par $\frac{n}{n_s}$, où $n_s$ est le nombre d'observations possédant la modalité $s$, et $m$ est le nombre total de modalités des $p_2 = \sum_{g=1}^{G} p_2^{(g)}$ variables qualitatives.

### 4.2.1 Algorithme de MFAmix

L'algorithme de MFAmix fonctionne de la manière suivante :

**Etape 1 : Construction de la matrice des poids des variables.**

1. Pour $g = 1, \ldots, G$, on récupère la première valeur propre $\lambda_1^{(g)}$ issue de PCAmix appliquée sur le groupe $g$ (représenté par la matrice $\mathbf{X}^{(g)}$).

2. On construit la matrice diagonale $\mathbf{P}$, de dimension $p_1 + m \times p_1 + m$, contenant les poids supplémentaires (nécessaires à l'AFM) des colonnes de $\mathbf{Z}$. Ainsi $\mathbf{P} = diag\left(\frac{1}{\lambda_1^{(t_k)}}\right)$, où $t_k \in \{1, \ldots, g, \ldots, G\}$ correspond au groupe de la $k$-ème colonne de $\mathbf{Z}$. On construit ensuite la matrice diagonale $\mathbf{M}^* = \mathbf{MP}$ prenant en compte les poids classiques des variables (contenus comme dans PCAmix dans la matrice $\mathbf{M}$) et les poids supplémentaires des variables, nécessaires pour l'AFM, contenus dans la matrice $\mathbf{P}$. Finalement, cette matrice $\mathbf{M}^*$ sera utilisée dans la GSVD.

**Etape 2 : Obtention des coordonnées factorielles.**

1. La GSVD de $\mathbf{Z}$ avec les métriques $\mathbf{N}$ et $\mathbf{M}^* = \mathbf{MP}$ donne la décomposition suivante
$$\mathbf{Z} = \mathbf{U}_{\text{mfa}}\mathbf{\Lambda}_{\text{mfa}}\mathbf{V}_{\text{mfa}}^t,$$
comme défini à l'équation (2.2.1).
2. La matrice $\mathbf{F}_{\text{mfa}}$ des coordonnées factorielles des lignes de $\mathbf{Z}$ est obtenue comme suit :
$$\mathbf{F}_{\text{mfa}} = \mathbf{U}_{\text{mfa}}\mathbf{\Lambda}_{\text{mfa}}. \tag{4.2.1}$$
3. La matrice $\mathbf{A}^*_{\text{mfa}}$ des coordonnées factorielles des colonnes de $\mathbf{Z}$ est obtenue de la manière suivante :
$$\mathbf{A}^*_{\text{mfa}} = \mathbf{MV}_{\text{mfa}}\mathbf{\Lambda}_{\text{mfa}}. \tag{4.2.2}$$

### 4.2.2 Correspondance avec les sorties classiques de PCAmix

Nous avons vu que la méthode MFAmix est équivalente à la méthode PCAmix appliquée sur des variables précédemment repondérées. Cependant en raison de cette repondération, les contributions des variables quantitatives et des modalités se calculent différemment.
La contribution absolue $c_{jk,\text{mfa}}$ d'une variable quantitative $j$ à la variance de la composante principale $k$ se calcule comme suit :
$$c_{jk,\text{mfa}} = \frac{1}{\lambda_1^{(t_j)}} a^{*2}_{jk,\text{mfa}}, \tag{4.2.3}$$
et la contribution absolue $c_{sk,\text{mfa}}$ d'une modalité $s$ à la variance de la composante principale $k$ se calcule comme suit :
$$c_{sk,\text{mfa}} = \frac{1}{\lambda_1^{(t_s)}} \frac{n_s}{n} a^{*2}_{sk,\text{mfa}}, \tag{4.2.4}$$
où $t_j$ (resp. $t_s$) $\in \{1, \ldots, G\}$ correspond au groupe de la $j$-ème variable quantitative (resp. de la variable qualitative contenant la $s$-ème modalité).

**Remarque :** Les contributions et les cosinus carrés des observations se calculent de la même façon que dans PCAmix. C'est également le cas pour les cosinus carrés des variables quantitatives et des modalités. Les contributions des variables qualitatives sont toujours égales à la somme des contributions des modalités qu'elles possèdent. Pour plus de détails, le lecteur peut se reporter à la Section 2.3.2.

#### 4.2.2.1 Calcul des "squared loadings" issus de MFAmix

Nous avons vu à la Section 2.3.2 la notion de "squared loading" qui permet de représenter sur un même plan factoriel les variables quantitatives et les variables qualitatives. Le "squared loading" est une mesure $\in [0,1]$ qui nous renseigne sur la liaison entre une variable et une composante principale (ou axe factoriel). On rappelle que le "squared loading" entre une variable quantitative et un axe factoriel est égal à la corrélation au carré entre la variable et l'axe. D'autre part, le "squared loading" entre une variable qualitative et un axe est égal au rapport de corrélation entre la variable et l'axe. Ceci reste valable dans MFAmix. Cependant, dans PCAmix le "squared loading" entre une variable (qualitative ou qualitative) et un axe factoriel était directement égal à sa contribution, cela n'est plus le cas dans MFAmix. Dans MFAmix, le "squared loading" $sql(j,k)$ entre une variable $j$ (quantitative ou qualitative) et un axe factoriel $k$ est égal à $\lambda_1^{(t_j)} c_{jk,\text{mfa}}$. D'après les équations (4.2.3) et (4.2.4), on peut réécrire les "squared loadings" comme suit :

$$\begin{aligned} sql_{\text{mfa}}(j,k) &= \lambda_1^{(t_j)} c_{jk,\text{mfa}} \\ &= \begin{cases} a^{*2}_{jk,\text{mfa}} & \text{si la variable } j \text{ est quantitative,} \\ \sum_{s\in\mathcal{M}_j} \frac{n_s}{n} a^{*2}_{sk,\text{mfa}} & \text{si la variable } j \text{ est qualitative,} \end{cases} \end{aligned}$$

où $\mathcal{M}_j$ est l'ensemble des modalités de la variable qualitative $j$.

#### 4.2.2.2 Coefficients des combinaisons linéaires associées aux composantes principales

Comme dans PCAmix, les composantes principales de MFAMix peuvent s'écrire comme une combinaison linéaire des vecteurs $\mathbf{z}_1, \ldots, \mathbf{z}_{p_1+m}$ qui sont les vecteurs colonnes de $\mathbf{Z}$. Cette écriture sous la forme d'une combinaison linéaire peut, par exemple, servir à calculer les coordonnées factorielles de nouvelles observations sur un axe factoriel. Cette combinaison linéaire s'écrit sous la forme suivante :

$$\mathbf{f}_{k,\text{mfa}} = \mathbf{Z}\mathbf{M}^*\mathbf{v}_{k,\text{mfa}} = \sum_{j=1}^{p_1} \frac{1}{\lambda_1^{(t_j)}} v_{jk,\text{mfa}} \mathbf{z}_j + \sum_{s=p_1+1}^{p_1+m} \frac{1}{\lambda_1^{(t_s)}} \frac{n}{n_s} v_{sk,\text{mfa}} \mathbf{z}_s.$$

On réécrit $\mathbf{f}_k$ comme la somme d'une combinaison linéaire des variables quantitatives $\mathbf{x}_j$, des indicatrices des modalités $\mathbf{x}_s$ et d'une constante $\beta_0$ :

$$\mathbf{f}_{k,\text{mfa}} = \beta_0 + \sum_{j=1}^{p_1} \beta_j \mathbf{x}_j + \sum_{s=p_1+1}^{p_1+m} \beta_s \mathbf{x}_s \tag{4.2.5}$$

où les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_{p_1+m}$ sont les colonnes de $\mathbf{X} = (\mathbf{X}_1|\mathbf{G})$. Et les coefficients $\beta_0$, $\beta_j$ et $\beta_s$ sont donnés par :

$$\begin{aligned}
\beta_0 &= -\sum_{l=1}^{p_1} v_{lk,\text{mfa}} \frac{\bar{\mathbf{x}}_l}{\sigma_l} \frac{1}{\lambda_1^{(t_l)}} - \sum_{l=p_1+1}^{p_1+m} v_{lk,\text{mfa}} \frac{1}{\lambda_1^{(t_l)}} \bar{\mathbf{x}}_l \\
\beta_j &= v_{jk,\text{mfa}} \frac{1}{\sigma_j} \frac{1}{\lambda_1^{(t_j)}}, \text{ pour } j = 1, \ldots, p_1 \\
\beta_s &= v_{sk,\text{mfa}} \frac{n}{n_s} \frac{1}{\lambda_1^{(t_s)}}, \text{ pour } s = p_1 + 1, \ldots, p_1 + m
\end{aligned}$$

avec $\bar{\mathbf{x}}_l$ et $\sigma_l$ la moyenne empirique et respectivement l'écart type de la colonne $\mathbf{x}_l$.

### 4.2.3 Sorties spécifiques à l'analyse factorielle multiple

Du fait de la structure en groupes de variables, certaines sorties graphiques et numériques sont spécifiques à l'AFM. Nous allons voir dans cette section comment les calculer dans le cadre de MFAmix. Comme pour la méthode PCAmix, chaque sortie sera accompagnée du code R nécessaire à son obtention. On suppose ici que le résultat de la fonction `MFAmix` est stocké dans l'objet `res.mfamix`.

#### 4.2.3.1 Sorties relatives aux groupes de variables

**Contributions des groupes de variables.** Afin d'observer les relations entre les groupes de variables et les axes factoriels, on représente les groupes en fonction de leurs contributions partielles aux axes factoriels. On définit la contribution absolue d'un groupe $g$ comme la somme des contributions absolues des variables qu'il contient. La contribution partielle d'un groupe $g$ à un axe factoriel $k$ se calcule donc comme suit :

$$cr(\mathbf{X}^{(g)}, k) = \frac{1}{\lambda_k} \sum_{j=1}^{p^{(g)}} c_{jk,\text{mfa}}$$

où $p^{(g)}$ est le nombre de variables (quantitatives et qualitatives) du groupe $g$, $c_{jk,\text{mfa}}$ est la contribution absolue de la variable $j$ (quantitative ou qualitative) à l'axe $k$ et $\lambda_k$ la valeur propre associée à l'axe $k$. Les contributions partielles ainsi que leurs représentations sont obtenues avec les lignes de code suivantes :

```
res.mfamix$groups$contrib.pct
plot(res.mfamix, choice="groups")
```

**Mesures de liaison entre les groupes.** Afin de mesurer les liaisons entre les groupes de variables et ainsi voir s'ils apportent la même information, nous définissons deux mesures de liaisons entre deux groupes de variables représentés par les matrices $\mathbf{X}^{(a)}$ et $\mathbf{X}^{(b)}$.

Tout d'abord le coefficient $Lg$ est défini comme suit :

$$Lg(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) = \frac{1}{n^2}\frac{1}{\lambda_1^{(a)}}\frac{1}{\lambda_1^{(b)}}\, Tr\left(\mathbf{Z}^{(a)t}\mathbf{Z}^{(b)}\mathbf{Z}^{(b)t}\mathbf{Z}^{(a)}\right),$$

où $\lambda_1^{(a)}$ (resp. $\lambda_1^{(b)}$) est la première valeur propre issue de PCAmix du groupe $\mathbf{X}^{(a)}$ (resp. $\mathbf{X}^{(b)}$) et $\mathbf{Z}^{(a)}$ (resp. $\mathbf{Z}^{(b)}$) est la matrice $\mathbf{X}^{(a)}$ (resp. $\mathbf{X}^{(b)}$) recodée comme dans PCAmix (voir Section 2.3.1) puis centrée et réduite.

Le coefficient $RV$ entre deux groupes de variables peut être vu comme une standardisation du coefficient $Lg$. En effet, le coefficient $Lg$ est dépendant de la taille des groupes et est donc difficilement interprétable. Le coefficient $RV$ est compris entre 0 et 1, il vaut 0 lorsque toutes les variables du groupe $\mathbf{X}^{(a)}$ sont orthogonales à toutes les variables du groupe $\mathbf{X}^{(b)}$ et il vaut 1 lorsque les groupes sont homothétiques. Le coefficient $RV$ entre deux groupes de variables est défini de la manière suivante :

$$\begin{aligned} RV(\mathbf{X}^{(a)}, \mathbf{X}^{(b)}) &= \frac{Lg(\mathbf{X}^{(a)}, \mathbf{X}^{(b)})}{\sqrt{Lg(\mathbf{X}^{(a)}, \mathbf{X}^{(a)})Lg(\mathbf{X}^{(b)}, \mathbf{X}^{(b)})}} \\ &= \frac{Tr\left(\mathbf{Z}^{(a)t}\mathbf{Z}^{(b)}\mathbf{Z}^{(b)t}\mathbf{Z}^{(a)}\right)}{\sqrt{Tr\left(\mathbf{Z}^{(a)t}\mathbf{Z}^{(a)}\mathbf{Z}^{(a)t}\mathbf{Z}^{(a)}\right)Tr\left(\mathbf{Z}^{(b)t}\mathbf{Z}^{(b)}\mathbf{Z}^{(b)t}\mathbf{Z}^{(b)}\right)}} \end{aligned}$$

Les coefficients $Lg$ et $RV$ sont des sorties de la fonction `MFAmix` et s'obtiennent avec les lignes de code suivantes :

```
res.mfamix$groups$Lg
res.mfamix$groups$RV
```

#### 4.2.3.2 Projection des axes partiels des analyses séparées

Le cercle des corrélations permet l'observation des liens entre les variables quantitatives mais également des liens entre les variables et les composantes principales de MFAmix, appelées composantes globales. Afin de faciliter la lecture du cercle des corrélations lorsque celui ci comporte un grand nombre de variables, il peut être intéressant de représenter les composantes principales des analyses séparées, appelées axes partiels ou composantes partielles, sur le cercle des corrélations de MFAmix. La coordonnée du l-ème axe partiel $\mathbf{f}_l^{(g)}$ du groupe $g$ sur l'axe factoriel de MFAmix $\mathbf{f}_k$ est égale à

$r(\mathbf{f}_l^{(g)}, \mathbf{f}_k)$, où $r$ désigne la corrélation de Pearson. Ainsi, la représentation de ces axes partiels permet d'observer leurs corrélations avec les composantes globales de MFAmix et donc de voir quels groupes de variables ont le plus contribué à leur construction. On peut également comparer les directions de dispersions d'inerties des analyses séparées et donc conclure à l'existence d'un potentiel facteur commun aux différents groupes. Les coordonnées des axes partiels et leurs représentations graphiques sont obtenues avec les lignes de code suivantes :

```
res$partial.axes$coord
plot(res, choice="axes")
```

#### 4.2.3.3 Projection des observations partielles

La représentation du nuage initial des $n$ observations de $\mathbb{R}^p$ est appelé nuage global. Le nuage global est représenté grâce aux coordonnées factorielles des observations contenues dans la matrice $\mathbf{F}_{\text{mfa}}$ définie à l'équation (4.2.1). On souhaite également représenter simultanément les nuages des projections des observations pour chaque groupe de variables. Pour cela on construit la matrice $\tilde{\mathbf{Z}}^{(g)}$ de dimension $n \times p$ des observations partielles du groupe $g$. Cette matrice est construite en conservant uniquement les valeurs des observations sur le groupe $g$ et en complétant le reste de la matrice par des zéros, comme suit :

$$\tilde{\mathbf{Z}}^{(g)} = \left(\mathrm{O}_{n\times p^{(1)}} | \dots | \mathrm{O}_{n\times p^{(g-1)}} | \mathbf{Z}^{(g)} | \mathrm{O}_{n\times p^{(g+1)}} | \dots | \mathrm{O}_{n\times p^{(G)}}\right),$$

où $\mathrm{O}_{n\times p^{(g)}}$ est la matrice nulle de dimension $n \times p^{(g)}$.

Ensuite on obtient les coordonnées factorielles des observations partielles en projetant les lignes de $\widetilde{\mathbf{X}}^{(g)}$ sur les axes de MFAmix. Cependant, la projection est légèrement différente de la projection d'observations supplémentaires classiques car, chaque coordonnée projetée est multipliée par $G$, le nombre de groupes. Ainsi chaque coordonnée factorielle globale est au barycentre des coordonnées factorielles des observations partielles. Les coordonnées factorielles $\widetilde{\mathbf{F}}^{(g)}$ des observations partielles du groupe $g$ sont définies comme suit :

$$\widetilde{\mathbf{F}}^{(g)} = G \times \widetilde{\mathbf{Z}}^{(g)} \mathbf{M}^* \mathbf{V}.$$

Les projections des nuages d'observations décrits par chaque groupe de variables sont appelées nuages partiels. Leur analyse permet d'observer quelle est l'influence des différents groupes quant à la position d'une observation sur un axe factoriel. Les coordonnées partielles des observations ainsi que leurs représentations graphiques sont obtenues à l'aide des lignes de code suivantes :

```
res$ind.partial$coord
plot(res, choice="ind", partial="all")
```

### 4.2.4 Illustration de la méthode MFAmix à l'aide du package `PCAmixdata`

Cette section illustre la méthode MFAmix à l'aide du package `PCAmixdata`. Le but ici n'est pas la construction d'indicateurs composites de qualité de vie mais simplement d'interpréter rapidement les différentes sorties de MFAmix. La méthode est utilisée sur les données présentes dans l'objet `gironde` du package et présenté à la Section 3.2.4. Nous utilisons ici les quatres groupes de variables disponibles, relatifs à quatre dimensions de la qualité de vie : l'emploi (groupe quantitatif `employment`, le logement (groupe mixte `housing`), l'accès à différents services (groupe qualitatif `services`) et l'environnement (groupe quantitatif `environment`). On rappelle que la description des quatre jeux de données est donnée en Annexe B. Les Figures 4.1 et 4.2 contiennent les sorties graphiques de MFAmix. Cet exemple n'étant qu'illustratif, nous nous limitons à l'examen du premier plan factoriel (axes 1 et 2). L'extrait de code ci-dessous détaille comment charger les données et comment utiliser la fonction `MFAmix`.

```
## chargement de la liste gironde contenant les differents dataframe
data(gironde)
## rassemblement des quatre datataframe dans un seul
dat<-cbind(gironde$employment[1:200, ], gironde$housing[1:200, ],
           gironde$services[1:200, ], gironde$environment[1:200, ])
## vecteur indiquant l'appartenance de chaque variable a chacun
## des groupes. Les 9 premieres au groupes 1, les 5 suivantes au
   groupe 2, ...
class.var<-c(rep(1,9),rep(2,5),rep(3,9),rep(4,4))
## vecteur des nom des groupes et des couleurs des groupes
names <- c("employment", "housing", "services", "environment")
color.group<-c("red", "black", "blue", "green")
## on lance la methode MFAmix
res<-MFAmix(data=dat, groups=class.var,
            name.groups=names, rename.level=TRUE, graph=FALSE)
```

**Cercle des corrélations des variables quantitatives.** La Figure 4.1(a) représente le cercle des corrélations des variables quantitatives colorées en fonction de leur appartenance aux différents groupes. Nous avons choisi d'afficher sur ce graphique les variables qui contribuent à plus de 50% de la variance associée au plan factoriel (1-2). On remarque que l'axe 1 est corrélé positivement à la densité de population (`density`) ainsi qu'au pourcentage de bâtiments sur la commune (`buildings`), ces deux variables étant corrélées négativement au pourcentage de propriétaires de leur logement (`owners`). L'information apportée par le second axe factoriel concerne principalement le groupe relatif à l'emploi (appelé `employment` et représenté en rouge). En effet le taux d'emploi (`employrate`), le revenu (`income`) et la part d'emplois intermédiaires et qualifiés (`middleempl` et `managers`) sont corrélés positivement avec cet axe, alors que le pourcentage de retraités (`retired`) est corrélé négativement avec l'axe 2.

Le cercle des corrélations présenté ici a été obtenu grâce au code suivant :

```
plot(res,choice="cor", coloring.var="groups",
     lim.contrib.plot = 0.5, col.groups=color.group,cex=0.8,main="(a)
        Correlation circle")
```

**Coordonnées factorielles des modalités.** La Figure 4.1(b) représente les coordonnées factorielles des modalités des variables qualitatives. Chaque modalité est colorée en fonction de son appartenance aux différents groupes. Nous avons choisi de représenter uniquement les modalités dont la qualité de représentation (cosinus carré) sur le premier plan factoriel est supérieure à 0.5. On remarque que l'axe 1 représente un gradient en terme de nombre de services sur la commune. En effet des valeurs négatives sur l'axe 1 indiquent une présence relativement faible de services, alors que des valeurs élevées sur l'axe 1 indiquent une présence assez abondante de services au sein de la commune. Le graphique des coordonnées des modalités est obtenu à l'aide du code suivant :

```
plot(res, choice="levels", coloring.var="groups",
     xlim=c(-2,2.5), cex=0.8, col.groups=color.group,
     lim.cos2.plot=0.5, main="(b) Levels")
```

**Coordonnées factorielles des observations.** La Figure 4.1(c) représente les coordonnées factorielles des observations. Sur cette figure, nous avons représenté uniquement les observations dont la qualité de représentation (cosinus carré) sur le premier plan factoriel est supérieure à 0.5. De plus chaque observation est colorée en fonction de ses valeurs sur la variable `postoffice`. On remarque que l'axe 1 est discriminant pour cette variable. En effet les communes ayant de fortes valeurs sur l'axe 1 sont majoritairement celles ayant un bureau de poste (en rouge sur le graphique). Inversement les communes avec de faibles valeurs sur l'axe 1 sont celles n'ayant pas de bureau de poste (colorées en noir). Le graphique a été obtenu avec le code suivant :

```
plot(res ,choice="ind" ,coloring.ind=dat$postoffice, cex=0.5,
     col.groups=color.group, lim.cos2.plot=0.5, posleg="topright",
     main="(c) Observations")
```

**"Squared loadings" de toutes les variables.** La Figure 4.1(d) représente l'ensemble des variables en fonction de leurs "squared loadings". On observe que les variables du groupe relatif aux services sont fortement liées à l'axe 1 de MFAmix. L'axe 2 est principalement lié aux variables du groupe `employment` relatif à l'emploi. Ce graphique est obtenu à l'aide du code suivant :

```
plot(res, choice="sqload", coloring.var="groups", cex=0.8,
     col.groups=color.group, posleg="topright", main="(d) All
        variables")
```

**Coordonnées factorielles des observations partielles.** La Figure 4.2(a) représente les coordonnées partielles des communes de Bassens et d'Arcachon. Ces deux communes ont des coordonnées globales (point central) très proches sur le graphique. Cependant, on remarque que leurs coordonnées partielles associées à certains groupes sont relativement éloignées. Par exemple la coordonnée partielle associée au groupe `housing` pour la commune d'Arcachon est nettement plus faible sur l'axe 2 que celle associée à la ville de Bassens. Au vu du cercle des corrélations présenté à la Figure 4.1(a) on observe que la variable `primaryres` est corrélée positivement à l'axe 2. Cette variable indique le pourcentage de résidences principales parmi les logements totaux sur la commune. On en déduit donc que, même si les communes de Bassens et d'Arcachon ont un profil assez similaire, la commune d'Arcachon a un pourcentage de résidences principales assez inférieur à celui de Bassens. Le graphique de ces deux observations partielles s'obtient avec le code suivant :

```
plot(res, choice="ind", partial=c("ARCACHON","BASSENS"),
     lim.cos2.plot=0.5, col.groups=color.group, cex=0.5,
     main="(a) Partial observations")
```

**Cercle des corrélations des axes partiels.** La Figure 4.2(b) représente les deux premiers axes factoriels (appelés axes partiels) de chaque analyse de groupe séparée. Ce graphique permet une représentation plus synthétique des relations entre les variables et les composantes principales. Dans un premier temps, ce graphique permet de voir quels groupes de variables sont les plus liés aux composantes globales de MFAmix. Puis, on peut également interpréter chaque analyse séparée afin de comprendre quelles variables sont le plus liées aux composantes partielles de chaque groupe. Par exemple, on observe que le premier axe factoriel associé au groupe `employment` est fortement corrélé à l'axe 2 de MFAmix. La Figure 4.2(c) représente le cercle des corrélations de l'analyse séparée du groupe `employment`. On remarque que l'axe 1 est corrélé positivement au taux d'emploi, au revenu et au taux de professions intermédiaires. L'analyse des deux graphiques présentés ici permet de conclure que les variables précédemment citées sont corrélées positivement à l'axe 2 de MFAmix, chose que nous avions déjà observé sur la Figure 4.1(a).

```
#Plot des axes partiels
plot(res,choice="axes", nb.partial.axes=2, coloring.var="groups",
     col.groups=color.group, cex=0.5, main="(b) Partial axes")
#Plot du cercle des correlations du groupe employment
plot(res$separate.analyses$employment, choice="cor",
     main="(c) Correlation circle of the group 'employment'")
```

**Contributions des groupes.** La Figure 4.2(d) représente les groupes en fonction de leurs contributions relatives aux composantes principales. On rappelle que la contri-

bution d'un groupe est égale à la somme des contributions des variables du groupe. Ce graphique montre par exemple que les trois groupes `environment`, `housing` et `services` contribuent chacun à environ 30% de la variance de la première composante principale. Alors que le groupe `employment` contribue à environ 80% de la variance de la seconde composante principale. Ce graphique est obtenu avec le code suivant :

```
plot(res, choice="groups",main="(d) Contributions of the groups")
```

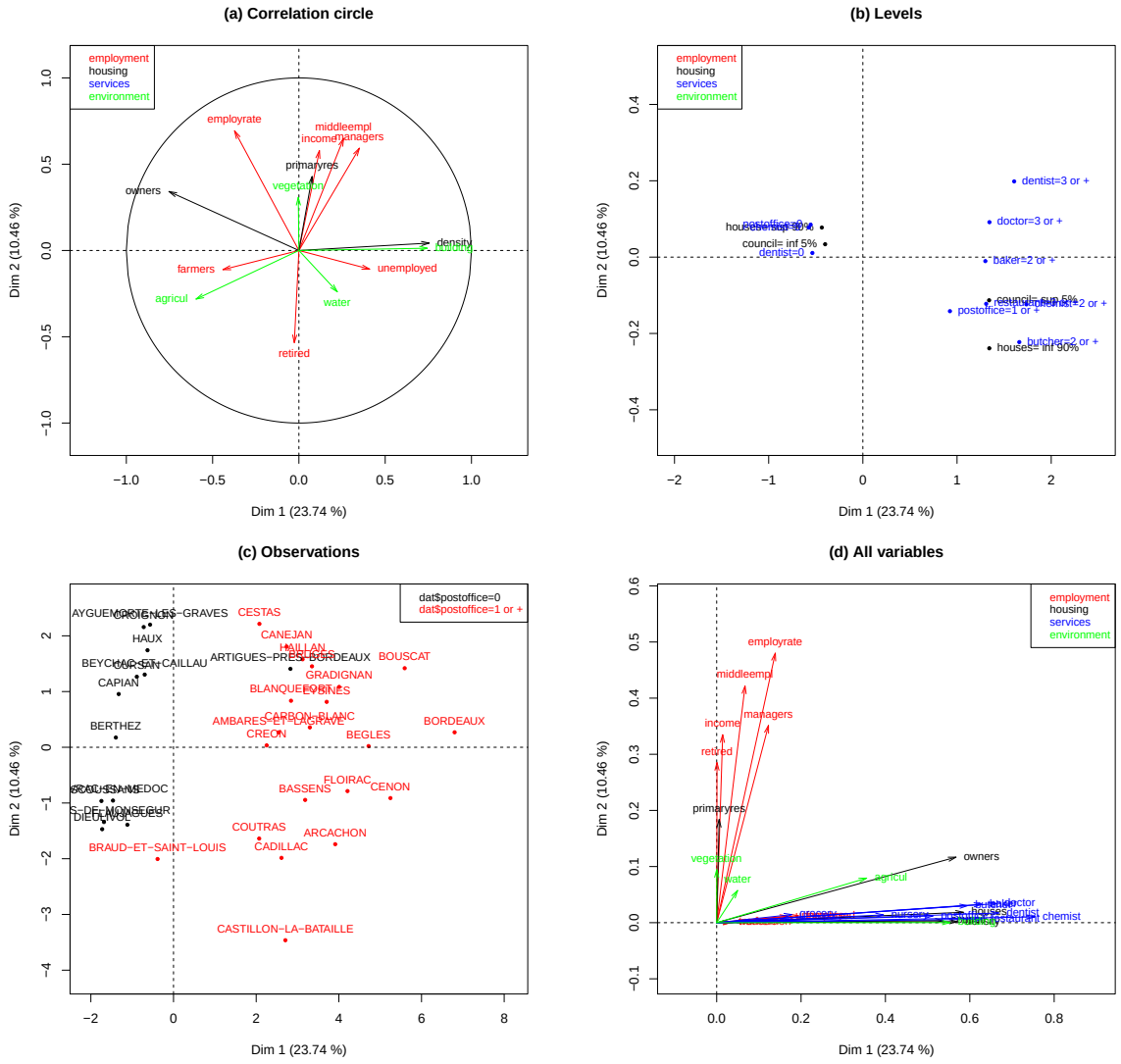



FIGURE 4.1 – (a) Cercle des corrélations. (b) Coordonnées factorielles des modalités. (c) Coordonnées factorielles des observations. (d) "Squared loadings" de toutes les variables.

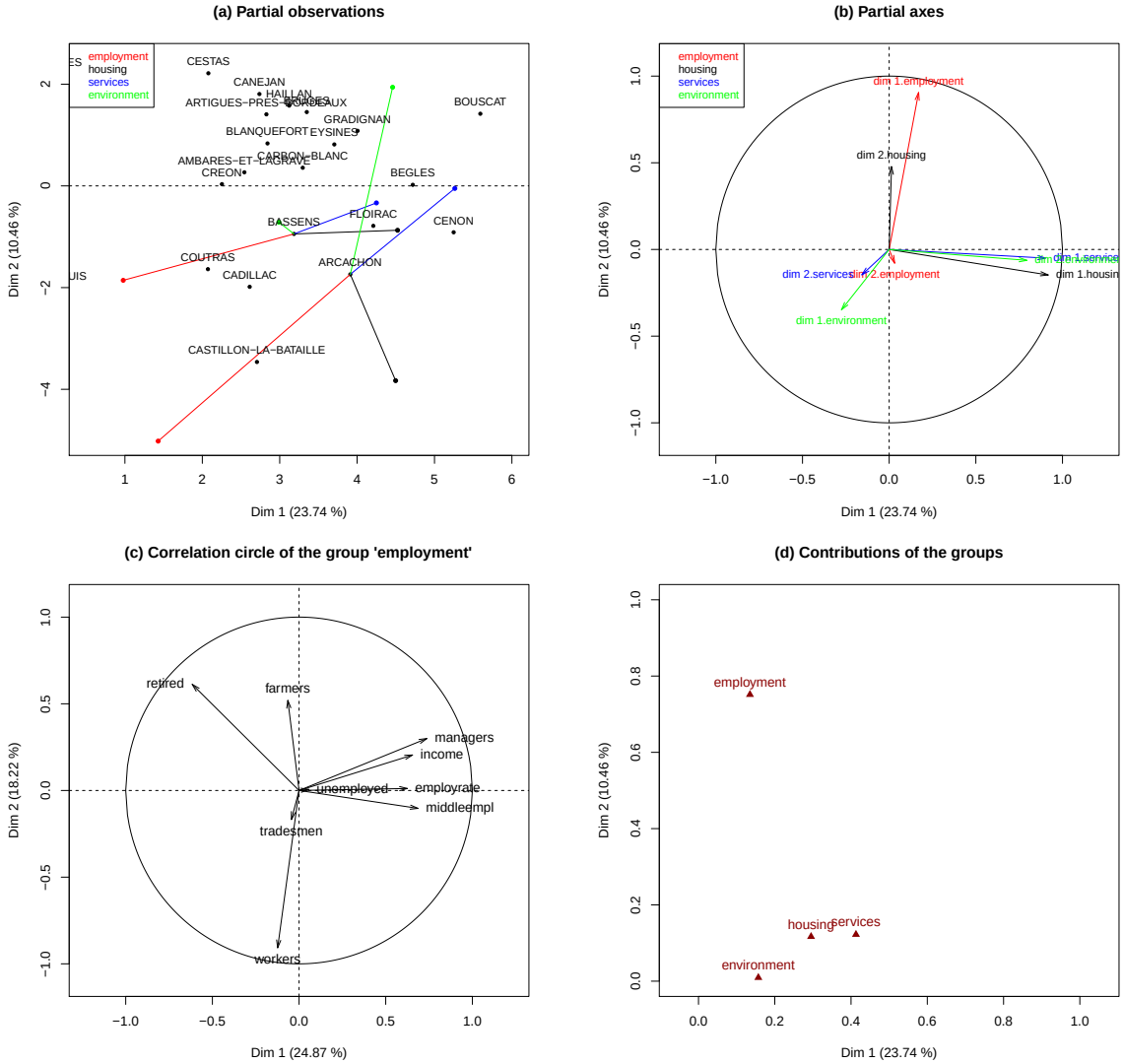



FIGURE 4.2 – (a) Coordonnées factorielles des observations partielles. (b) Cercle des corrélations des axes partiels. (c) Cercle des corrélations du groupe `employment`. (d) Contributions des groupes.

## 4.3 Sélection de variables au sein de MFAmix

La méthode MFAmix présentée à la Section 4.2 permet la construction de composantes principales orthogonales qui résument l'information apportée par l'ensemble des variables réparties en différents groupes. Nous verrons à la Section 4.4 comment ces composantes principales peuvent être vues comme des indicateurs de qualité de vie à l'échelle communale et comment les interpréter. Cependant ces composantes principales étant calculées sur l'ensemble des $p$ variables disponibles, les indicateurs obtenus sont donc des combinaisons linéaires des $p$ variables incluses dans l'analyse. Si le nombre $p$ est important, il peut être difficile de donner un sens aux indicateurs obtenus. L'intérêt de cette section est double. Tout d'abord nous proposons une méthode afin de choisir un nombre de composantes principales à interpréter, en se basant sur la notion de distances entre sous-espaces. Par la suite, une fois que le nombre de composantes principales à interpréter a été fixé, nous proposons une méthode permettant de sélectionner un nombre restreint $p^*$ de variables ($p^* < p$) à inclure dans l'analyse tout en obtenant des nouvelles composantes principales les plus corrélées possible aux composantes principales calculées sur l'ensemble des $p$ variables. Cela permettra de faciliter l'interprétation des composantes principales en perdant le moins d'information possible.

### 4.3.1 Choix du nombre de composantes principales de MFAmix à interpréter

Nous allons utiliser ici un critère de stabilité dans le but de choisir un nombre de composantes principales pertinent. Il existe différents critères de choix de dimension en analyse factorielle : voir par exemple Besse (1992) ou Josse and Husson (2012). Le critère que nous utilisons ici est celui initialement développé par Besse (1992) dans le cadre de l'ACP. Il sera réutilisé ici dans le cadre de MFAmix mais peut être appliqué de la même manière sur toutes les méthodes d'analyse factorielle basée sur des GSVD. Ce critère de stabilité utilise la notion de distance entre projecteurs et est construit comme une fonction de risque estimée par bootstrap des observations.
On rappelle ici quelques résultats et notations relatifs à MFAmix. La méthode MFAmix est basée sur la GSVD de $\mathbf{Z}$ (la matrice $n \times p$ des données brutes précédemment recodées) avec les métriques $\mathbf{N}$ contenant les poids des observations et $\mathbf{M}$ contenant les poids des variables. On a ainsi :

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t,$$

avec :

— $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ la matrice des valeurs singulières de $\mathbf{ZNZ}^t\mathbf{M}$ et $\mathbf{Z}^t\mathbf{NZM}$ où $r$ est le rang de $\mathbf{Z}$,

— $\mathbf{U}$ la matrice $n \times r$ des vecteurs propres de $\mathbf{ZMZ}^t\mathbf{N}$ et $\mathbf{U}^t\mathbf{DU} = \mathbb{I}_r$,

— $\mathbf{V}$ la matrice $p \times r$ des vecteurs propres de $\mathbf{Z}^t\mathbf{NZM}$ et $\mathbf{V}^t\mathbf{MV} = \mathbb{I}_r$.

On construit la matrice $\mathbf{F} = \mathbf{U\Lambda}$ des composantes principales de MFAmix.
On définit ensuite la matrice de projection $\widehat{\mathbf{P}_q} = \mathbf{V}_q\mathbf{V}_q^t\mathbf{M}$ pour $1 < q < r$, qui est la matrice de projection $\mathbf{M}$-orthogonale des lignes de $\mathbf{Z}$ sur $E_q = Im(\mathbf{V}_q)$, le sous-espace engendré par les $q$ premières colonnes de $\mathbf{V}$.

La fonction de perte reposant sur la distance euclidienne entre deux projecteurs orthogonaux est donnée par :

$$\begin{aligned}\mathcal{L}_q = \mathcal{Q}(E_q, \widehat{E_q}) &= \frac{1}{2}\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\|_2^2 \\ &= \frac{1}{2}Tr[(\mathbf{P}_q - \widehat{\mathbf{P}_q})(\mathbf{P}_q - \widehat{\mathbf{P}_q})^t] \\ &= q - Tr(\mathbf{P}_q\widehat{\mathbf{P}_q})\end{aligned}$$

Finalement, le risque est défini comme l'espérance de la fonction de perte :

$$R_q = E[\mathcal{L}_q].$$

L'idée est d'estimer $R_q$ par $\widehat{R_q}$ à l'aide d'une méthode de rééchantillonage par bootstrap. On définit d'abord la fonction de perte $\mathcal{L}_q^b$ qui est la fonction de perte associée à un échantillon bootstrap $b$.

$$\mathcal{L}_q^b = q - Tr(\widehat{P_q^{*b}}\widehat{P_q}),$$

où :

— $\widehat{P_q^{*b}}$ est la matrice de projection obtenue sur le $b$-ème échantillon bootstrap,

— $\widehat{P_q}$ est la matrice de projection obtenue sur l'échantillon de départ.

Ainsi, on peut définir l'estimateur bootstrap $\widehat{R_{Bq}}$ de $R_q$ de la manière suivante :

$$\widehat{R_{Bq}} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_q^b = q - Tr(\widehat{P_q^{*(.)}}\widehat{P_q}),$$

où :

— $B$ est le nombre d'échantillons bootstrap,

— $\widehat{P_q^{*(.)}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{P_q^{*b}}$.

On cherche à choisir une valeur de $q$ telle que les sous-espaces engendrés par les échantillons bootstrap soient les plus proches possible du sous-espace de base $E_q = Im(\mathbf{V}_q)$. Pour cela, l'estimateur bootstrap $\widehat{R_{Bq}}$ doit prendre des valeurs proches de zéro et les valeurs des fonctions de pertes $\mathcal{L}_q^b$ associées à chaque échantillon bootstrap ne doivent pas être trop dispersées. Une faible dispersion de ces valeurs indique la stabilité du sous-espace correspondant.
Pour illustrer cette méthode, on va l'appliquer aux résultats de MFAmix présentés à la

Section 4.2.4. Ainsi pour chaque $q \in [\![1, 10]\!]$, on génère $B = 1000$ échantillons bootstrap et on calcule $\mathcal{L}_q^b$ pour $b \in [\![1, B]\!]$ puis on calcule $\widehat{R_{Bq}}$. La Figure 4.3 représente pour chaque valeur de $q$ un boxplot des $B$ valeurs des fonctions de perte $\mathcal{L}_q^b$. Pour chaque valeur de $q$, l'estimateur $\widehat{R_{Bq}}$ est également représenté en rouge. Au vu de ce graphique, on observe que les fonctions de pertes $\mathcal{L}_q^b$ sont très proches de zéro pour $q = 1$, cependant on considère que l'interprétation d'un seul axe factoriel n'est pas suffisante. On choisit donc d'interpréter $q = 3$ composantes principales car malgré la dispersion des fonctions de pertes $\mathcal{L}_q^b$ pour $q = 3$, l'estimateur $\widehat{R_{Bq}}$ prend une valeur assez proche de zéro.

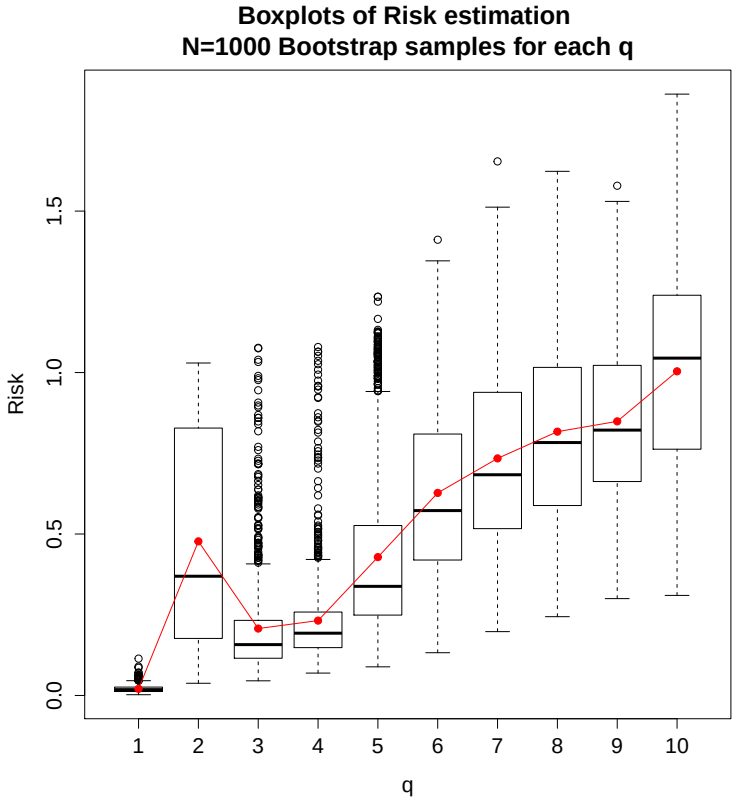



FIGURE 4.3 – Boxplot des fonctions de pertes $\mathcal{L}_q^b$ et estimateur $\widehat{R_{Bq}}$ (en rouge).

### 4.3.2 Sélection de variables à l'aide de la méthode "Closest Submodel Selection"

Une fois le nombre de composantes principales à interpréter fixé, nous allons chercher à obtenir des nouvelles composantes principales, calculées sur un nombre restreint de variables. Pour sélectionner les variables à introduire dans MFAmix, nous allons nous baser sur la méthode "Closest Submodel Selection" (CSS), voir Coudret et al. (2014), développée dans le cadre de la régression SIR. Le but de cette méthode est de sélectionner un sous-ensemble de $p^\star < p$ variables, tel que les $q$ ($q = 3$ dans notre exemple)

premières composantes principales de MFAmix calculées sur ce sous-ensemble de $p^\star$ variables soient le plus liées possible aux $q$ premières composantes principales calculées sur l'ensemble des $p$ variables. On introduit les notations suivantes :

- $\mathbf{F} = \mathbf{U}\mathbf{\Lambda}$ est la matrice $n \times q$ des composantes principales issues de MFAmix réalisée sur les $p$ variables. Et $P_F = \mathbf{F}\mathbf{F}^T\mathbf{N}$ est la matrice de projection $\mathbf{N}$-orthogonale sur $\mathbf{F}$.
- $\mathbf{F}^\star = \mathbf{U}^\star\mathbf{\Lambda}^\star$ est la matrice $n \times q$ des composantes principales issues de MFAmix réalisée sur les $p^\star < p$ variables. Et $P_{F^\star} = \mathbf{F}^\star\mathbf{F}^{\star T}\mathbf{N}$ est la matrice de projection $\mathbf{N}$-orthogonale sur $\mathbf{F}^\star$.

Comme expliqué précédemment, nous souhaitons obtenir des composantes principales $\mathbf{F}^\star$ les plus liées possible aux composantes principales de référence $\mathbf{F}$. Cela nécessite donc de définir une mesure de liaison entre deux groupes de composantes principales, cette mesure est définie comme suit :

$$\mathcal{D}(\mathbf{F}\mathbf{F}^\star) = \frac{1}{q} Tr(P_F P_{F^\star}). \tag{4.3.1}$$

**Mise en œuvre de la méthode CSS dans le cadre de MFAmix.** Nous présentons ici le déroulement de la méthode CSS de sélection de variables dans le cadre de MFAmix. L'idée de la méthode est de sélectionner plusieurs sous-ensembles de $p_0$ variables. Pour chaque sous-ensemble, on réalise MFAmix, puis on regarde la liaison entre les composantes principales obtenues ($\mathbf{F}^\star$) et les composantes principales de référence calculées sur toutes les variables ($\mathbf{F}$). Par la suite on ne sélectionne que les meilleurs sous-ensembles (au sens de la mesure de liaison entre les composantes principales). Les variables les plus présentes dans ces meilleurs sous-ensembles apparaissent alors comme les meilleures variables pour obtenir des composantes principales $\mathbf{F}^\star$ le plus liées à $\mathbf{F}$. L'algorithme fonctionne comme suit :

**Etape 0.** On fixe $p_0 \in [\![q, p[\![$ le nombre de variables de chacun des sous-ensembles à évaluer. On fixe $N_0$ le nombre de sous-ensembles à évaluer et $\zeta$ le pourcentage des meilleurs sous-ensembles à conserver.

**Etape 1.** Pour $a = 1, \ldots, N_0$, on répète :

- On sélectionne $p_0$ variables parmi $p$ et on construit la matrice $\mathbf{X}^{(a)}$ contenant les variables sélectionnées.
- On réalise MFAmix sur $\mathbf{X}^{(a)}$ et on calcule $\mathcal{D}(\mathbf{F}\mathbf{F}^{(a)})$.

**Etape 2.** On conserve les $N_1 = \zeta N_0$ "meilleurs" sous-ensembles, c'est à dire ceux dont la mesure de liaison entre leurs composantes principales et les composantes principales $\mathbf{F}$ de référence est la plus grande.

**Etape 3.** On compte le nombre de fois où chaque variable apparaît dans les $N_1$ meilleurs sous-ensembles de variables. Finalement, les variables apparaissant le plus souvent sont retenues pour effectuer MFAmix.

**Remarque :** En pratique, le choix du paramètre $p_0$ a peu d'influence sur les résultats. En effet, dès lors que le nombre $N_0$ de sous-ensembles à évaluer est très grand, les $p$ variables se retrouveront en moyenne dans un même nombre de sous-ensembles à évaluer.

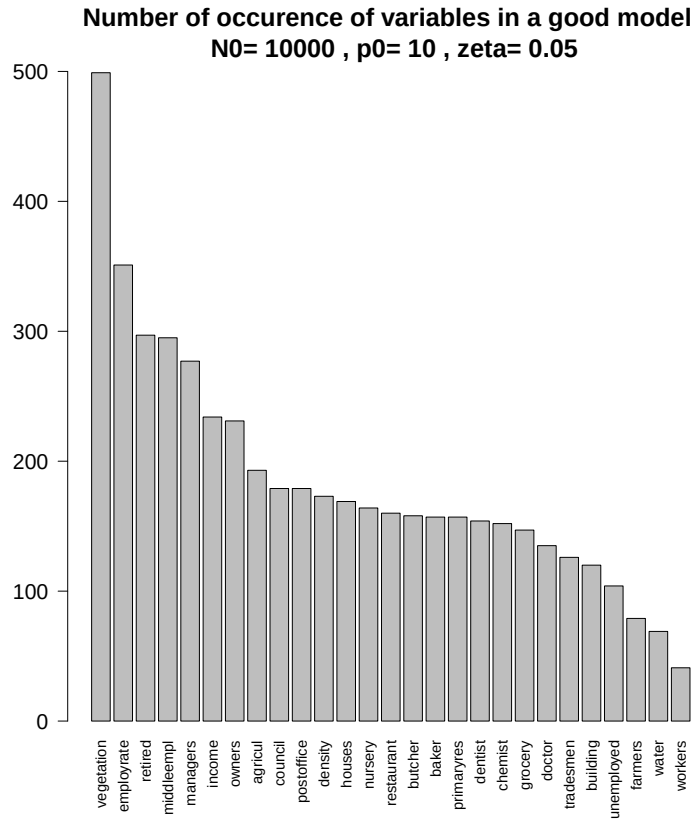



FIGURE 4.4 – Nombre d'apparition de chaque variable dans les meilleurs sous-ensembles.

Comme nous l'avons fait pour l'illustration de la méthode permettant de choisir un nombre de composantes principales à interpréter, nous allons illustrer la méthode CSS sur les résultats de MFAmix présentés à la Section 4.2.4. Comme vu précédemment on fixe le nombre de composantes principales à interpréter à $q = 3$. On rappelle que ces composantes principales ont été obtenues grâce à la méthode MFAmix réalisée sur les $p = 27$ variables du jeu de données `gironde` contenu dans le package `PCAmixdata`.
La méthode CSS a été réalisée sur $N_0 = 10000$ sous-ensembles de $p_0 = 10$ variables. Nous avons choisi de conserver $\zeta = 5\%$ des meilleurs sous-ensembles (c'est à dire $N_1 = 500$ "bons" sous-ensembles). La Figure 4.4 représente le nombre de fois où chaque variable est apparue dans un bon sous-ensemble. On remarque par exemple que la variable `vegetation` est présente dans tous les bons sous-ensembles. Par la suite, afin de sélectionner quelles variables introduire dans MFAmix, nous allons réaliser la méthode

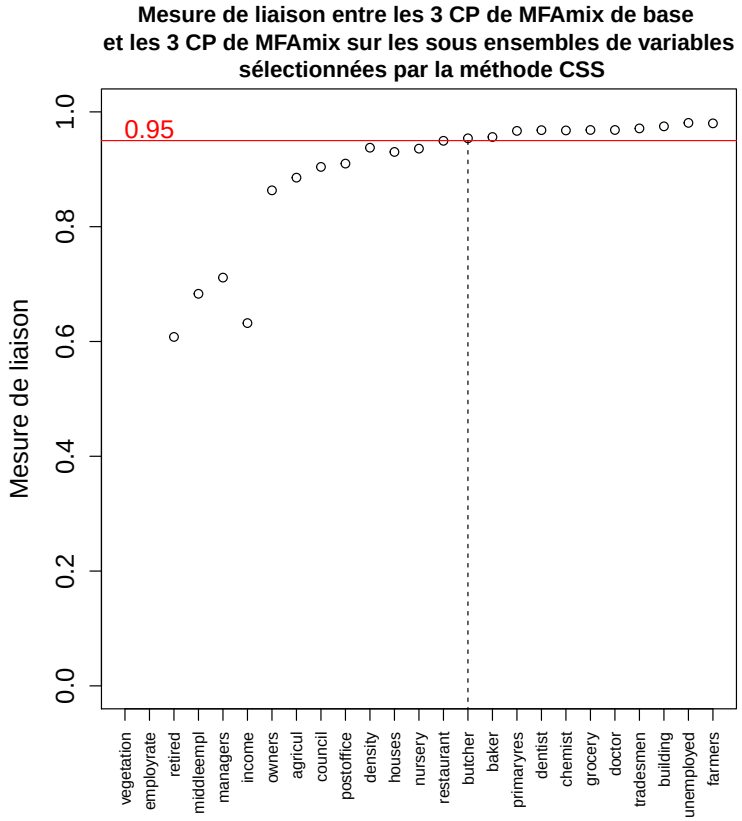



FIGURE 4.5 – Mesures de liaisons entre composantes principales en fonction du sous-ensemble de variables.

sur les trois premières variables qui apparaissent le plus dans les bons sous-ensembles (`vegetation`, `employrate` et `retired`) puis calculer la mesure de liaison entre les composantes principales obtenues et les composantes principales de référence (calculées sur toutes les variables). Ensuite, nous ferons de même avec les quatre premières variables (`vegetation`, `employrate`, `retired` et `middleempl`), puis les 5 et ainsi de suite jusqu'à introduire toutes les variables. A chaque fois que l'on rajoute une variable dans MFAmix on regarde la mesure de liaison entre les composantes principales obtenues et les composantes principales de référence. Finalement, on se fixe un seuil de mesure de liaison (par exemple 0.95) entre composantes principales à dépasser. Les variables retenues seront celles telles que lorsque l'on réalise MFAmix sur celles-ci, on obtient une mesure de liaison supérieure au seuil fixé. Ces résultats sont présentés à la Figure 4.5. On voit par exemple que la mesure de liaison entre les composantes principales de MFAmix réalisée sur les trois premières variables (`vegetation`, `employrate` et `retired`) est de l'ordre de 0.6. Si on inclut les 15 premières variables (de `vegetation` à `butcher`) on obtient une mesure de liaison supérieure à 0.95. Ces 15 variables permettent donc d'obtenir des composantes principales très corrélées (corrélation supérieure à 0.95) aux composantes principales de référence (calculées avec les $p = 27$ variables d'origine), ces corrélations prises deux à deux sont rassemblées dans la Table 4.1.

| | CP 1.15var | CP 2.15var | CP 3.15var |
|---|---|---|---|
| CP 1.ref | 0.98 | -0.03 | 0.09 |
| CP 2.ref | 0.02 | 0.98 | 0.09 |
| CP 3.ref | -0.11 | -0.08 | 0.96 |

TABLE 4.1 – Corrélations entre les composantes principales de référence calculées sur les $p = 27$ variables (CP .ref) et les composantes principales obtenues avec les $p^\star = 15$ variables (CP .15var).

## 4.4 Construction d'indicateurs composites de qualité de vie à l'aide de la méthode MFAmix

Cette section est dédiée à la construction d'indicateurs composites de qualité de vie à l'aide de la méthode MFAmix présentée précédemment. Comme nous l'avons vu, cette méthode permet de tenir compte de la structuration des données en groupes de variables. Les composantes principales de MFAmix seront interprétées en fonction des variables qui leur sont le plus corrélées. Ces composantes peuvent s'écrire comme des combinaisons linéaires de l'ensemble des données d'origine, elles peuvent donc être

considérées comme des indicateurs composites. Par la suite, nous verrons comment simplifier ces indicateurs grâce à la méthode CSS présentée à la Section 4.3.2. Les indicateurs seront construits sur un échantillon de communes liées aux territoires de l'eau.

### 4.4.1 Présentation de la zone d'étude, des données et de la méthodologie adoptée

**La zone littorale des SAGES.** Nous avons choisi dans cette application d'effectuer un diagnostic territorial de communes associées aux "territoires de l'eau". Afin de définir ces territoires associés à l'eau nous avons voulu éviter de choisir les communes selon un découpage administratif (communauté de communes, département, région,...) mais plutôt de suivre des délimitations hydrographiques. Ainsi, nous avons délimité notre zone d'étude à l'aide des Schémas d'Aménagement et de Gestion des Eaux (SAGE). Le SAGE est un document de planification de la gestion de l'eau à l'échelle d'une unité hydrographique cohérente (bassin versant, aquifère, etc.). Il fixe des objectifs généraux d'utilisation, de mise en valeur, de protection quantitative et qualitative de la ressource en eau et il doit être compatible avec le Schéma Directeur d'Aménagement et de Gestion des Eaux (SDAGE). Nous avons donc retenu les 5 SAGEs suivants pour définir notre zone d'étude (le nombre de communes par SAGE est donné à la Table 4.2) :

- **Estuaire de la Gironde et milieux associés**, qui comprend l'agglomération Bordelaise ;
- **Seudre**, dont le pôle urbain est Royan ;
- **Leyre et cours d'eaux côtiers**, qui englobe la communauté urbaine d'Arcachon ;
- **Étangs littoraux de Born et Buch** ;
- **Lacs Médocains**.

Nous avons également intégré la commune d'Arcachon qui ne fait partie d'aucun SAGE afin de préserver la continuité spatiale de la zone d'étude. Nous arrivons finalement à une zone de 303 communes des départements de la Gironde, des Landes et de la Charente-Maritime et deux régions administratives, l'Aquitaine et le Poitou-Charentes. Cet ensemble de communes sera appelé par la suite la zone des SAGEs. Cette zone d'étude inclut des territoires littoraux emblématiques de la région, le Bassin d'Arcachon ainsi que l'estuaire de la Gironde. À l'interface entre terre et mer, ces territoires sont pourvus de nombreuses aménités environnementales et patrimoniales produites par leur richesse écologique et écosystémique, qui participent à les rendre très attractifs. Alors que leurs ressources naturelles, halieutiques, culturelles et patrimoniales leur confèrent une richesse à préserver, elles ne les protègent pas de la périurbanisation liée

à leur proximité des pôles urbains de Bordeaux ou Royan, l'avancée du front urbain engendrant des flux imposés par la ville et subis par le monde rural. Ces territoires, confrontés à des enjeux divers, entre attractivité, vulnérabilité, adaptation au risque et protection, constituent ainsi un terrain d'étude privilégié.

| SAGE | Nombre de communes |
|---|---|
| Estuaire de la Gironde et milieux associés | 185 |
| Seudre | 67 |
| Leyre et cours d'eaux côtiers | 43 |
| Étangs littoraux de Born et Buch | 27 |
| Lacs Médocains | 13 |
| Hors SAGE | 1 |
| Total | 303 |

Note : Certaines communes font partie de plusieurs SAGEs

TABLE 4.2 – Répartition des communes étudiées par SAGE.

**Données utilisées.** Nous avons présenté à la Section 1.2.2 le type de données que nous allons utiliser pour la construction des indicateurs composites. Les 56 variables utilisées ont été mesurées en 2009 sur l'ensemble des communes de la zone des SAGEs. On remarque que les variables mesurées sont les mêmes utilisés dans le Chapitre 3 pour la création d'indicateurs à l'aide de la méthode hclustvar. Les variables sont structurées en 6 groupes relatifs à 6 dimensions de la qualité de vie. La description de chacune des variables est donnée en Annexe D.

**Méthodologie adoptée.** Dans un premier temps nous appliquons la méthode MFAmix sur les données présentées précédemment. Nous choisissons un nombre de composantes principales (indicateurs composites de qualité de vie) grâce à la méthode présentée à la Section 4.3.1. Une fois les indicateurs composites construits, nous essayons de leur donner un sens en regardant quelles variables initiales sont le plus liées aux différents indicateurs sans faire ici de la sélection de variables. Puis, à l'aide des indicateurs composites construits, nous réalisons une typologie des communes par classification ascendante hiérarchique (CAH) afin de mieux identifier les ressemblances entre les différentes communes en terme de qualité de vie. Par la suite, nous utilisons la méthode CSS de sélection de variables au sein de MFAmix présentée à la Section 4.3. Cette méthode permet d'obtenir de nouveaux indicateurs composites simplifiés (calculés sur un nombre restreint de variables) et donc plus simples à écrire en tant que combinaisons linéaires des variables initiales. Une autre méthodologie aurait pu être adoptée en

réalisant la typologie des communes sur les indicateurs composites simplifiés obtenus avec la méthode de sélection de variables. Cependant, nous avons choisi de réaliser la typologie sur les indicateurs de référence (calculés sur l'ensemble des variables d'origine) afin de conserver le plus possible d'information. La simplification des indicateurs est réalisée par la suite et de manière indépendante dans le but de faciliter leur écriture et leur lecture.

### 4.4.2 Résultats de MFAmix et indicateurs composites créés

Une fois la méthode MFAmix réalisée, la première étape de l'analyse consiste à choisir un nombre pertinent de composantes principales à interpréter. En tant que combinaisons linéaires de l'ensemble des variables d'origine, ces composantes principales seront nos indicateurs composites de qualité de vie. Le choix du nombre de composantes principales est effectué grâce à la méthode présentée à la Section 4.3.1. La Figure 4.6 présente les résultats obtenus. Le graphique suggère de choisir $q = 4$ composantes principales à interpréter. En effet, l'estimateur $\widehat{R_{Bq}}$ est relativement faible jusqu'à $q = 4$ puis, on observe un saut à partir de la cinquième composante principale. Par la suite, les graphiques et les interprétations associées concerneront uniquement les $q = 4$ premières composantes principales de MFAmix.

La Figure 4.7 contient quatre graphiques utilisés pour l'interprétation des composantes principales de MFAmix. Les graphiques (a) et (b) correspondent respectivement aux cercles des corrélations des variables quantitatives sur les plans (1-2) et (3-4). Seules les variables dont la qualité de représentation sur le plan factoriel considéré est supérieure à 0.5 sont représentées. Le graphique (c) représente les coordonnées factorielles des modalités des variables qualitatives sur le plan (1-2). Ici aussi, seules les modalités dont la qualité de représentation est supérieure à 0.5 sont représentées. Le graphique (d) représente les corrélations entre les axes partiels des analyses séparées et les composantes principales globales de MFAmix. En nous basant sur ces graphiques ainsi que sur les sorties numériques associées nous essayons de donner un sens aux quatre premières composantes principales de MFAmix.

— La première composante principale de MFAmix est corrélée négativement avec le pourcentage de résidences de type maison, ainsi qu'avec le pourcentage de logements occupés par leurs propriétaires (variables `RPTypeMai` et `RPOccProp`). Cette composante principale est corrélée positivement avec la densité de population et le pourcentage de territoires bâtis sur la commune (variables `Densite` et `Bati`). De plus, on remarque sur la Figure4.7(c) un gradient de présence de services. En effet des faibles valeurs de la première composante principale indiquent l'absence ou la faible présence de services. Alors que des valeurs importantes de la première composante principale indiquent une offre abondante de services.

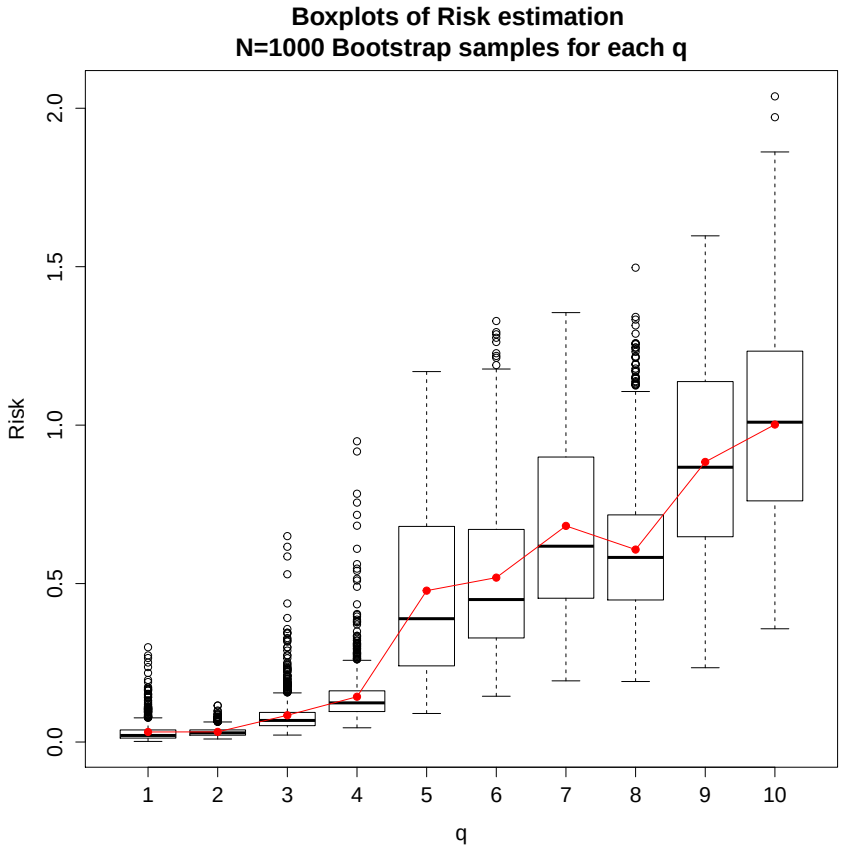



FIGURE 4.6 – Boxplot des fonctions de pertes $\mathcal{L}_q^b$ et estimateur $\widehat{R_{Bq}}$ (en rouge) sur la zone des SAGEs.

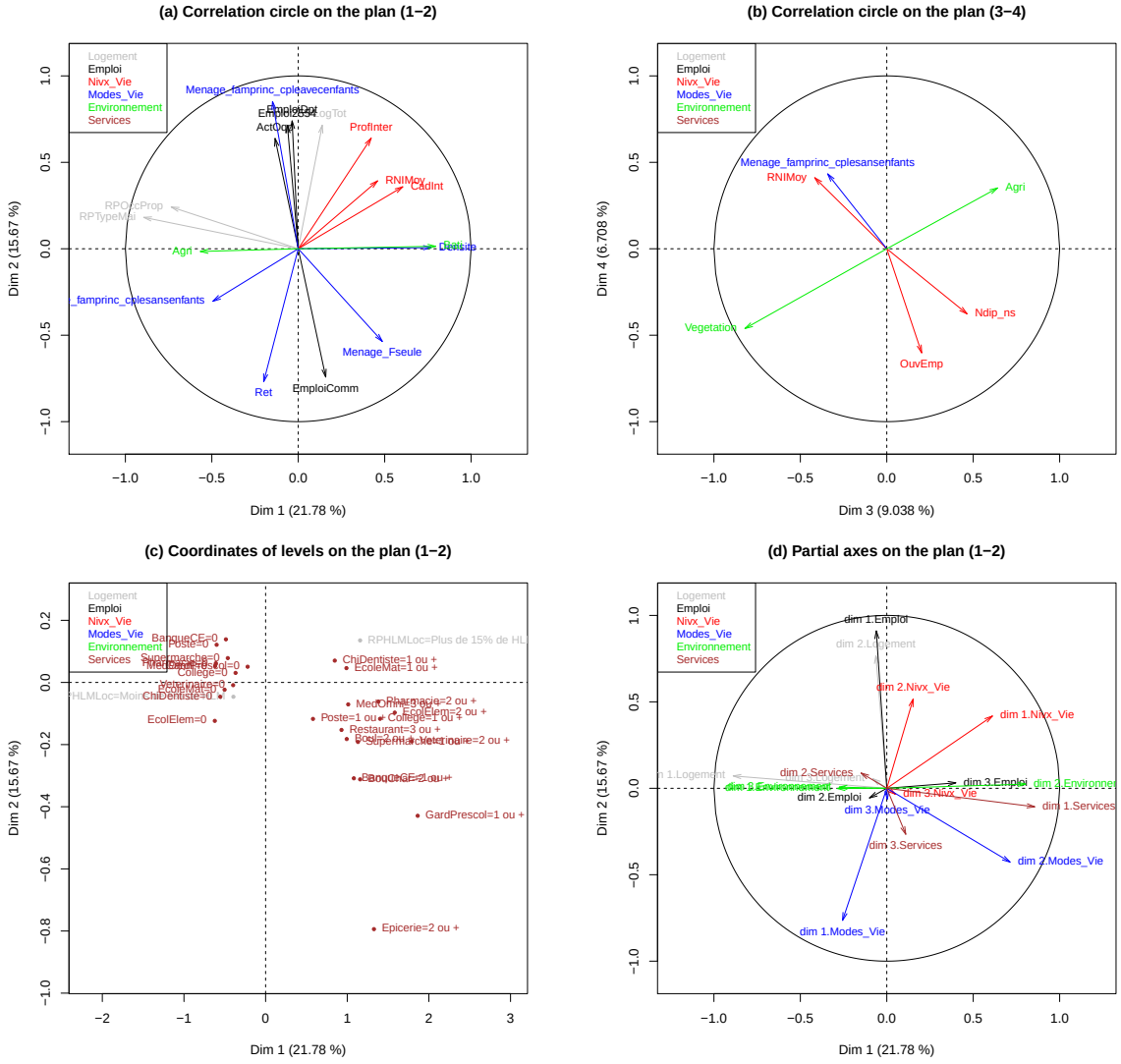



FIGURE 4.7 – (a, b) Cercle des corrélations des variables quantitatives sur les plans (1-2) et (3-4). (c) Coordonnées factorielles des modalités sur le plan (1-2). (d) Axes partiels des analyses séparées sur le plan (1-2).

— La seconde composante principale est corrélée négativement avec le pourcentage d'emplois au sein de la commune et le pourcentage de retraités (variables `EmploiComm` et `Ret`). Elle est également corrélée positivement au taux de logements principaux parmi les résidences principales (variable `RPLogTot`). Elle est aussi corrélée positivement au taux d'actifs occupés (variable `ActOqp`), au taux d'emploi chez les 25-54 ans (variable `Emploi2554`) et au pourcentage d'emplois à l'échelle du département (variable `EmploiDpt`). Pour finir, cette composante principale est également corrélée positivement avec le pourcentage de professions intermédiaires parmi tous les emplois de la commune (variable `ProfInter`) ainsi qu'avec le pourcentage de familles principales constituées d'un couple avec des enfants (variable `Menage_famprinc_cpleavecenfants`.

— La troisième composante principale est fortement liée à la dimension environnementale. En effet, cette composante principale est corrélée négativement au pourcentage de végétation sur la commune (variable `Vegetation`) et elle est corrélée positivement au pourcentage de territoires agricoles sur la commune (variable `Agri`).

— La quatrième composante principale est corrélée négativement avec la proportion d'emplois ouvriers (variable `OuvEmp`). Dans une moindre mesure, on note une corrélation négative avec la proportion de la population non diplômée (variable `Ndip_ns`) et une corrélation positive avec le revenu moyen des habitants (variable `RNIMoy`).

Cette interprétation des composantes principales permet de leur donner un nom en fonction des variables qui leur sont le plus corrélées. On utilisera par la suite ces composantes comme indicateurs composites. On peut également relier les valeurs de ces indicateurs (positives ou négatives) aux valeurs des variables initiales. Cela permet de connaître le profil d'une commune en fonction de ses scores sur les différentes indicateurs. On crée ainsi la Table 4.3 qui est une aide à la lecture des indicateurs composites obtenus sur la zone des SAGEs.

On peut par exemple connaître le profil de la commune d'Arcachon grâce à ces scores sur les différentes composantes principales. Les scores sont obtenus à l'aide du code R suivant :

```
res$ind$coord["ARCACHON", ]
#     dim 1      dim 2      dim 3      dim 4
#      5.01      -3.97     -0.304       1.38
```

Ainsi, la commune d'Arcachon est une commune relativement urbaine avec une densité de population élevée et un accès aux services importants (forte valeur sur la première composante principale). Elle possède une proportion importante d'habitants travaillant sur la commune et un nombre assez élevé de logements secondaires (faible valeur sur la seconde composante principale). Le score d'Arcachon sur la troisième

| **Indicateur composite** | **Valeurs négatives** | **Valeurs positives** |
|---|---|---|
| IC 1 : Urbanisation | Beaucoup de logements de type maison | Peu de logements de type maison |
| | Proportion importante de propriétaires | Peu de propriétaires |
| | Faible densité de population | Densité de population élevée |
| | Peu de bâtiments | Proportion importante de bâtiments |
| | Peu de services | Beaucoup de services |
| IC 2 : Offre d'emplois et structure des ménages | Proportion importante d'emplois à l'échelle de la commune | Peu d'emplois à l'échelle de la commune |
| | Peu d'emplois à l'échelle du département | Nombre importants d'emplois à l'échelle du département |
| | Peu de professions intermédiaires | Pourcentage important de professions intermédiaires |
| | Proportion importante de logements secondaires | Proportion importante de logements principaux |
| | Proportion importante de retraités | Faible proportion de retraités |
| | Faible taux d'emplois | Taux d'emplois élevé |
| | Peu de familles avec enfants | Proportion importante de familles avec enfants |
| IC 3 : Environnement | Proportion importante de territoires végétalisés | Faible proportion de territoires végétalisés |
| | Faible proportion de territoires agricoles | Proportion importante de territoires agricoles |
| IC 4 : Qualification de l'emploi | Proportion importante d'emplois ouvriers | Peu d'emplois ouvriers |
| | Proportion importante de personnes non diplomées | Peu de personnes non diplomées |

Table 4.3 – Lecture des indicateurs composites sur la zone des SAGEs.

composante principale étant proche de 0, on en déduit que cette commune possède peu de territoires agricoles et de territoires végétalisés. Le score positif sur la quatrième composante principale indique que les habitants de cette commune sont plutôt diplômés et que la commune dispose de peu d'emplois ouvriers.

### 4.4.3 Typologie des observations sur les indicateurs composites créés

Nous utilisons ici les 4 indicateurs composites construits précédemment afin d'établir une typologie des communes. Pour cela, nous réalisons une classification ascendante hiérarchique avec critère de Ward sur les valeurs prises par les communes sur les indicateurs. Cette classification d'observations a pour but de rassembler dans une même classe les communes qui se ressemblent. Nous avons retenu une partition en 6 classes de communes, ce choix du nombre de classes a été effectué au vu du dendrogramme (non représenté ici) mais également afin d'avoir des profils de communes distincts et interprétables. L'interprétation des classes créées est réalisée à l'aide de la Table 4.4 qui donne les valeurs moyennes des indicateurs composites pour les différentes classes de communes.

A l'aide de la Table 4.4 et de la Table 4.3 d'aide à la lecture des indicateurs composites, on peut interpréter les classes de communes de la manière suivante :

| **Indicateur composite** | **Classe de communes** | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| | *n=83* | *n=46* | *n=66* | *n=46* | *n=19* | *n=43* |
| IC 1 : Urbanisation | **-0.75** | **1.14** | **-1.48** | -0.31 | **4.70** | **0.76** |
| IC 2 : Emplois et ménages | **0.66** | **-1.91** | -0.44 | **-0.63** | 0.50 | **1.90** |
| IC 3 : Environnement | **0.88** | -0.43 | **0.34** | **-1.12** | **1.32** | **-1.15** |
| IC 4 : Qualification de l'emploi | **-0.29** | 0.19 | **0.84** | **-1.02** | 0.24 | 0.05 |

En gras : valeurs significativement différentes de la moyenne globale de l'indicateur (par construction la moyenne globale est nulle) ; p-value inférieure à $10^{-5}$.

TABLE 4.4 – Valeurs moyennes des indicateurs composites sur les cinq classes de communes de la zone des SAGEs.

- **La classe 1** contient 83 communes principalement rurales avec une forte proportion de territoires agricoles. Ces communes sont habitées par une proportion élevée de familles avec enfants et on y retrouve assez peu de retraités. Les logements sont principalement de type maisons et occupés par les propriétaires. Les emplois se situent principalement à l'échelle du département et se sont en majorité des emplois peu qualifiés et des emplois ouvriers.
- **La classe 2** rassemble 46 communes qui se trouvent pour la plupart sur le littoral ou le long de l'estuaire. Ces communes ont un accès correct aux services et on

y trouve une proportion élevée de retraités. L'emploi se situe principalement à l'échelle de la commune. Et on y trouve également une proportion supérieure à la moyenne de logements secondaires.

- **La classe 3** rassemble 66 communes à dominante agricole. En effet, ces communes possèdent un pourcentage élevé de terres agricoles, l'accès aux services est relativement restreint et la densité de population est inférieure à la moyenne. Les logements sont en majorité des maisons occupées par leurs propriétaires. Les emplois principaux sont des emplois agricoles. On note également une proportion de retraités supérieure à la moyenne. Ces communes sont situées en majorité au nord de l'estuaire.
- **La classe 4** contient 46 communes fortement végétalisées. Ces communes correspondent principalement à des territoires forestiers. Ce sont des communes avec une faible densité de population, une faible proportion de bâtiments et un accès aux services restreint. L'emploi se situe principalement à l'échelle de la commune.
- **La classe 5** contient 19 communes appartenant au pôle urbain de Bordeaux. Ce sont des communes denses, avec une forte proportion de bâtiments, peu de logements de type maison et un très bon accès aux services. Les logements sont principalement de type locatif.
- **La classe 6** rassemble 43 communes situées autour du pôle urbain de Bordeaux. Ce sont des communes relativement urbaines. On y trouve une proportion assez importante de familles avec enfants. L'emploi se situe principalement à l'échelle du département et les habitants de ces communes possèdent des emplois intermédiaires ou qualifiés avec un revenu moyen supérieur à la moyenne.

La Figure 4.8 représente la carte de la typologie obtenue.

### 4.4.4 Construction d'indicateurs simplifiés à l'aide de la méthode CSS

La méthode MFAmix réalisée sur la zone des SAGEs a permis de retenir quatre indicateurs composites de qualité de vie (les quatre premières composantes principales de MFAmix). Ces indicateurs composites (appelés par la suite composantes principales de référence) sont des combinaisons linéaires des $p = 45$ variables incluses dans l'analyse (soit 75 coefficients en incluant les modalités des variables qualitatives). Afin de faciliter l'écriture de ces indicateurs , nous allons chercher à construire de nouveaux indicateurs composites qui seront d'une part des combinaisons linéaires d'un nombre restreint de variables et d'autres part fortement corrélés aux indicateurs composites de référence (calculés sur les $p = 45$ variables d'origine). Nous allons pour cela utiliser la

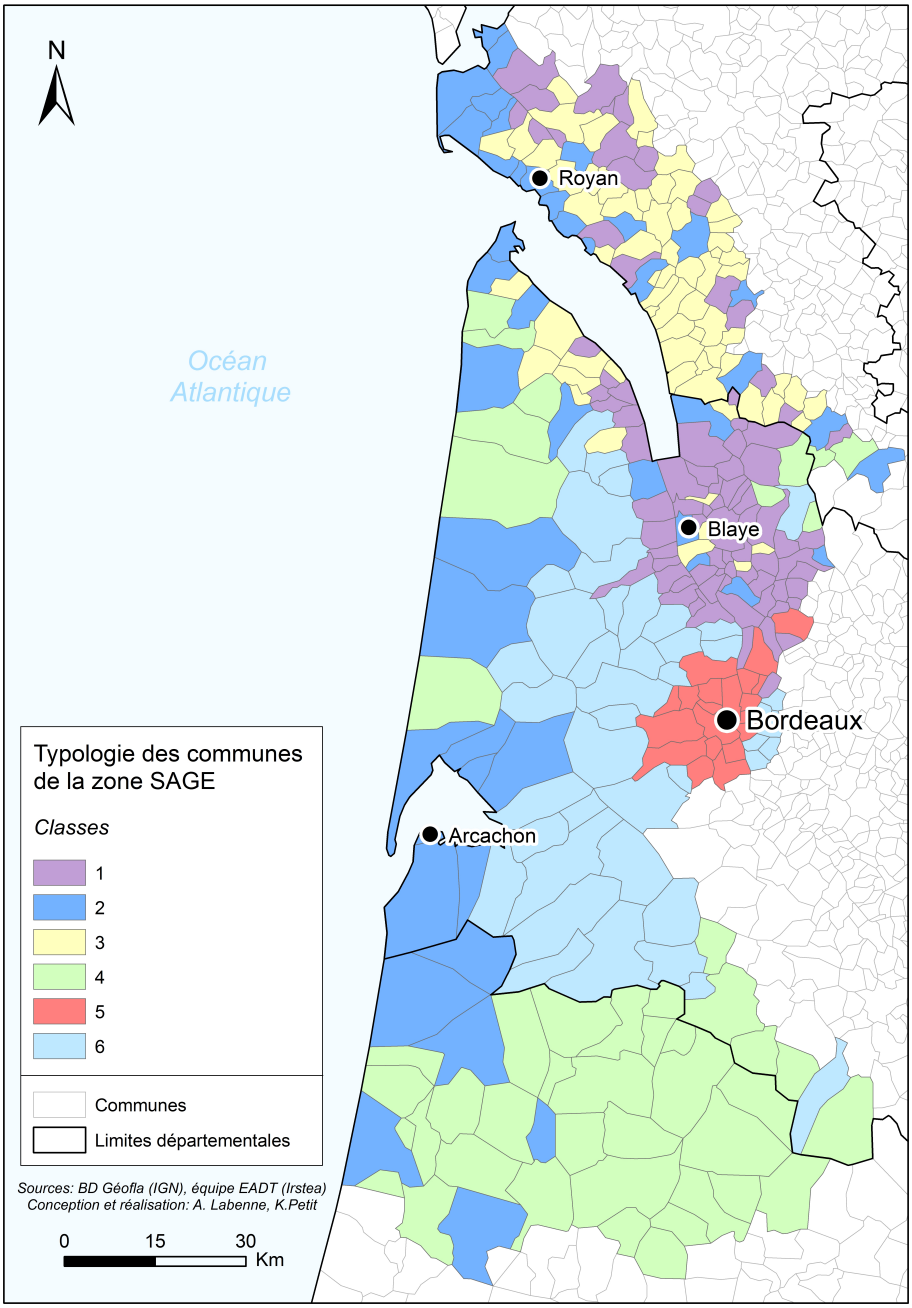



FIGURE 4.8 – Carte de la typologie des communes de la zone des SAGEs réalisée avec les indicateurs composites.

méthode CSS présentée à la Section 4.3.2.

La méthode CSS a été réalisée sur $N_0 = 10000$ sous-ensembles de $p_0 = 15$ variables. Nous avons choisi de conserver $\zeta = 5\%$ des meilleurs sous-ensembles (c'est à dire $N_1 = 500$ "bons" sous-ensembles). La Figure 4.9 représente le nombre de fois où chaque variable est apparue dans un bon sous-ensemble. On remarque par exemple que la variable `vegetation` est la plus présente dans les bons sous-ensembles. Par la suite, afin de sélectionner quelles variables introduire dans MFAmix, nous allons appliquer MFAmix sur les quatre premières variables qui apparaissent le plus dans les bons sous-ensembles (`Vegetation`, `Menage_famprinc_cplesansenfants`, `Ret` et `OuvEmp`) puis calculer la mesure de liaison entre les quatre composantes principales obtenues et les quatre composantes principales de référence. Ensuite, nous ferons de même avec les cinq premières variables, puis les 6 et ainsi de suite jusqu'à introduire toutes les variables. A chaque fois que l'on rajoute une variable dans MFAmix on regarde la mesure de liaison entre les composantes principales obtenues et les composantes principales de référence. Finalement, on se fixe un seuil de mesure de liaison (ici 0.95) entre composantes principales à dépasser. Les variables retenues seront celles telles que lorsque l'on réalise MFAmix sur celles-ci, on obtient une mesure de liaison supérieure au seuil fixé.

Ces résultats sont présentés à la Figure 4.10. On voit par exemple que la mesure de liaison entre les composantes principales de MFAmix réalisée sur les quatre premières variables (`Vegetation`, `Menage_famprinc_cplesansenfants`, `Ret` et `OuvEmp`) est environ égale à 0.6. Si on inclut les 20 premières variables (de `Vegetation` à `Supermarche`) on obtient une mesure de liaison supérieure à 0.95. Ces $p^\star = 20$ variables permettent donc d'obtenir des composantes principales très corrélées (corrélation supérieure à 0.95) aux composantes principales de référence, ces corrélations prises deux à deux sont rassemblées dans la Table 4.5.

| | CP 1 .20var | CP 2 .20var | CP 3 .20var | CP 4 .20var |
|---|---|---|---|---|
| CP 1.ref | 0.95 | 0.25 | 0.01 | -0.03 |
| CP 2.ref | -0.26 | 0.96 | -0.02 | 0.05 |
| CP 3.ref | 0.00 | 0.01 | 0.97 | 0.10 |
| CP 4.ref | 0.05 | -0.05 | -0.07 | 0.97 |

TABLE 4.5 – Corrélations entre les composantes principales de référence calculées sur les $p = 45$ variables (CP .ref) et les composantes principales obtenues avec les $p^\star = 20$ variables (CP .20var) sur la zone des SAGEs.

Les composantes principales simplifiées étant fortement corrélées positivement aux composantes principales de référence, nous pouvons utiliser la même Table 4.3 d'aide à la lecture de ces composantes. Cependant, afin de montrer que les composantes princi-

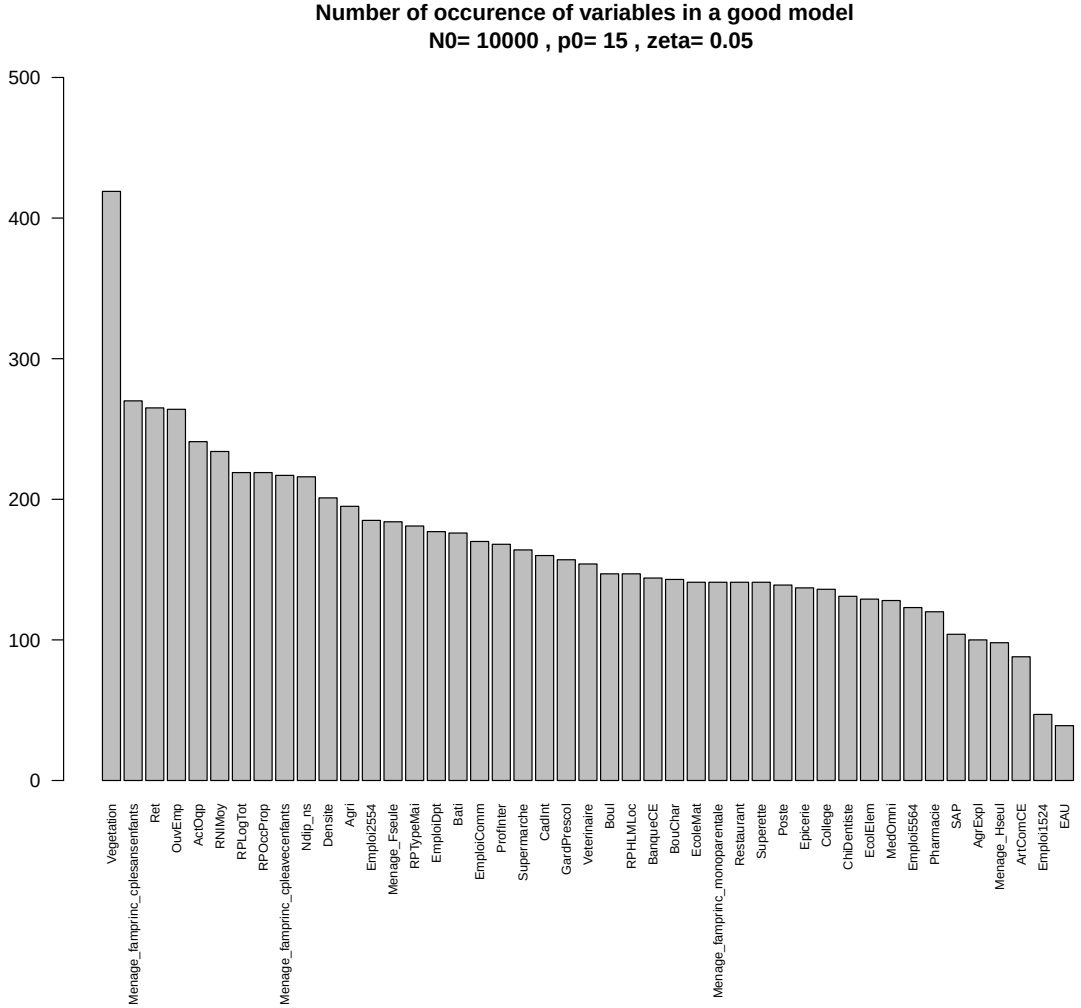



FIGURE 4.9 – Nombre d'apparitions de chaque variable dans les meilleurs sous-ensembles sur la zone des SAGEs.

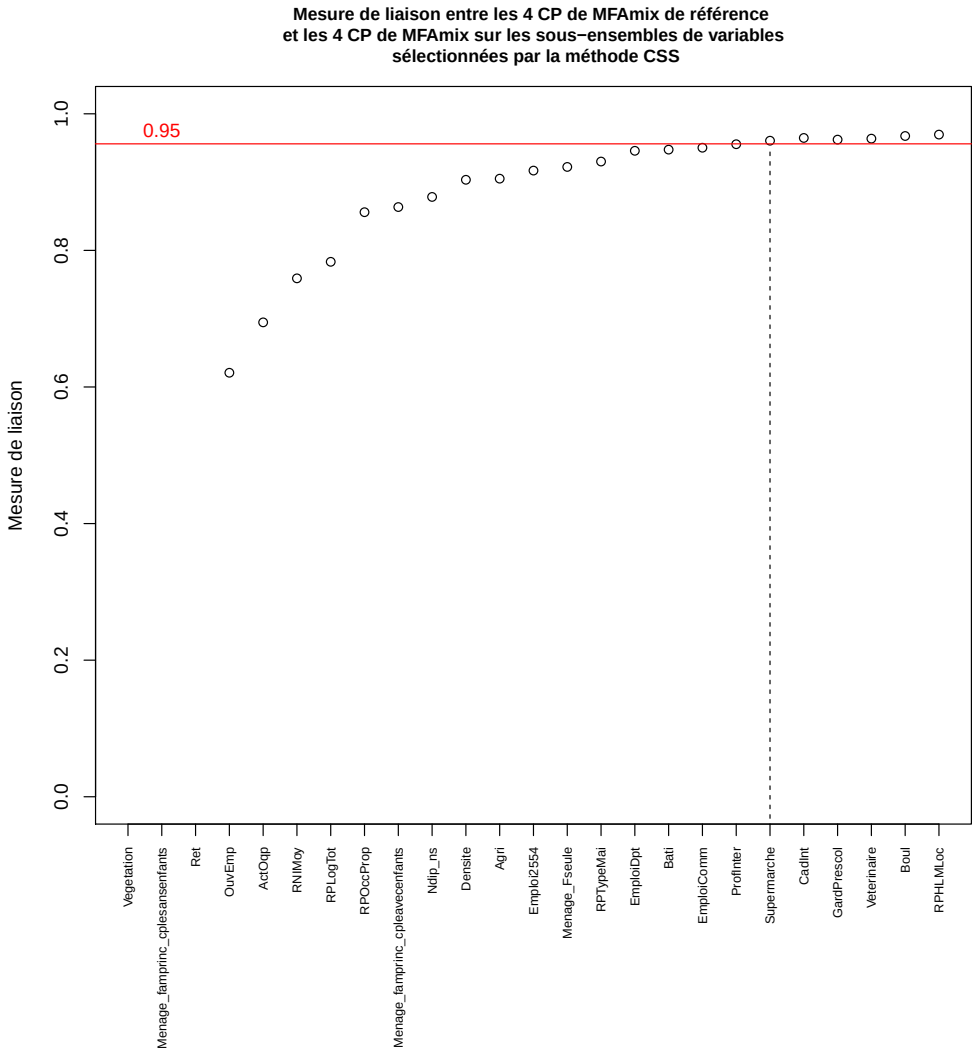



FIGURE 4.10 – Mesures de liaisons entre composantes principales en fonction du sous-ensemble de variables, sur la zone des SAGEs.

pales simplifiées sont quand même plus simples à écrire que les composantes principales de référence, nous avons choisi d'écrire la première composante principale (de référence et simplifiée) explicitement en tant que combinaison linéaire des variables initiales. Le calcul des coefficients associés à chaque variable (ou modalité d'une variable qualitative) est donné à la Section 4.2.2.2. La Table 4.6 donne pour la première composante principale de référence et pour la première composante principale simplifiée les coefficients associés à chaque variable (ou modalité d'une variable qualitative) nécessaires au calcul de la dite composante. On voit rapidement que la composante principale simplifiée est nettement plus facile à écrire que la composante principale de référence car c'est une combinaison linéaire de seulement 20 variables. Cette écriture simplifiée permet une meilleure diffusion et une meilleure compréhension de l'indicateur composite. L'examen des coefficients présenté à la Table 4.6 révèle que sur les 21 coefficients associés à la composante principale simplifiée, 19 coefficients ont le même signe sur la composante principale de référence. Deux coefficients ont un signe opposé, il s'agit des coefficients associés aux variables `Ret` et `RPLogTot`. Cela peut sembler problématique, cependant, si on regarde l'interprétation des composantes principales à la Table 4.3, on remarque que ces deux variables n'entrent pas en compte dans la caractérisation de la première composante principale. Ces deux variables interviennent par contre dans l'interprétation de la seconde composante principale, voir Table 4.3. Lorsque l'on regarde les coefficients associés à ces deux variables, ils ont le même signe sur la seconde composante principale de référence et sur la seconde composante principale simplifiée. Ces changements de signes occasionnels sur certaines composantes principales ont donc peu d'impact. Ces différences de signes sur quelques variables s'expliquent par le fait que la méthode CSS a permis de simplifier un sous-espace de dimension 4 et non pas quatre sous-espaces de dimension 1 pris séparément.

## 4.5 Conclusion

Ce chapitre a permis de présenter la méthode MFAmix. Cette méthode d'analyse factorielle permet de prendre en compte la structure des données en groupe de variables. De plus, contrairement à l'analyse factorielle multiple "classique", la méthode MFAmix permet la mixité des données (variables quantitatives et variables qualitatives) au sein d'un même groupe. Dans un premier temps nous avons détaillé l'utilisation et l'interprétation des résultats de MFAmix à l'aide du package `PCAmixdata`. Ensuite, nous avons vu comment cette méthode peut être utile pour construire des indicateurs composites de qualité de vie qui sont les composantes principales issues de l'analyse. Par la suite, une typologie des communes a été réalisée à l'aide des indicateurs composites obtenus. Cette typologie a permis de mieux comprendre les profils des communes étudiées.

| Variable | Coeff de la CP de référence ($\times 10^{-2}$) | Coeff de la CP simplifiée ($\times 10^{-2}$) | Variable | Coeff de la CP de référence ($\times 10^{-2}$) | Coeff de la CP simplifiée ($\times 10^{-2}$) |
|---|---|---|---|---|---|
| CONSTANTE | 371.94 | 775.80 | GardPrescol=0 | -1.24 | - |
| Vegetation | 0.11 | 0.10 | EcoleMat=1 ou + | 5.61 | - |
| Menage_famprinc_cplesansenfants | -1.96 | -1.69 | MedOmni=0 | -3.50 | - |
| Ret | -0.56 | 0.08 | Epicerie=2 ou + | 7.52 | - |
| OuvEmp | -0.51 | -1.11 | BouChar=2 ou + | 6.56 | - |
| ActOqp | -0.93 | -2.23 | BanqueCE=0 | -2.75 | - |
| RNIMoy | 0.01 | 0.01 | Menage_Hseul | 0.45 | - |
| RPLogTot | 0.21 | -0.16 | GardPrescol=1 ou + | 10.56 | - |
| RPOccProp | -1.77 | -2.50 | RPHLMLoc=Plus de 15% de HLM | 28.98 | - |
| Menage_famprinc_cpleavecenfants | -0.42 | -1.16 | Pharmacie=0 | -3.45 | - |
| Ndip_ns | -1.36 | -1.68 | Boul=2 ou + | 5.63 | - |
| Densite | 0.03 | 0.03 | Superette=1 ou + | 4.81 | - |
| Agri | -0.57 | -0.57 | EcolElem=1 | -0.13 | - |
| Emploi2554 | -0.28 | -1.13 | EcolElem=0 | -3.52 | - |
| Menage_Fseule | 2.08 | 3.06 | College=0 | -2.08 | - |
| RPTypeMai | -1.88 | -2.62 | ChiDentiste=1 ou + | 4.80 | - |
| EmploiDpt | -0.04 | -0.29 | BouChar=0 | -2.63 | - |
| Bati | 8.57 | 9.0 | College=1 ou + | 7.94 | - |
| EmploiComm | 0.26 | 0.61 | EcolElem=2 ou + | 8.96 | - |
| ProfInter | 1.83 | 1.54 | Boul=0 | -3.40 | - |
| Supermarche=0 | -2.62 | -30.67 | Pharmacie=2 ou + | 7.83 | - |
| Supermarche=1 ou + | 6.41 | 74.94 | EcoleMat=0 | -2.84 | - |
| EAU | 0.44 | - | MedOmni=3 ou + | 5.73 | - |
| Poste=1 ou + | 3.29 | - | Boul=1 | -2.00 | - |
| ChiDentiste=0 | -3.15 | - | Epicerie=1 | -0.74 | - |
| RPHLMLoc=Moins de 5% de HLM | -9.87 | - | Superette=0 | -0.92 | - |
| SAP | 2.21 | - | Restaurant=0 | -3.84 | - |
| Emploi1524 | -0.64 | - | Veterinaire=2 ou + | 10.15 | - |
| BanqueCE=1 ou + | 6.13 | - | Pharmacie=1 | 0.20 | - |
| Emploi5564 | 0.80 | - | MedOmni=1 ou 2 | -0.84 | - |
| BouChar=1 | -0.19 | - | Veterinaire=0 | -2.25 | - |
| Menage_famprinc_monoparentale | 1.97 | - | Restaurant=1 | -2.28 | - |
| AgrExpl | -2.52 | - | Epicerie=0 | -1.57 | - |
| ArtComCE | -1.32 | - | Restaurant=3 ou + | 5.27 | - |
| CadInt | 3.45 | - | Veterinaire=1 | 4.10 | - |
| Poste=0 | -3.40 | - | Restaurant=2 | -1.03 | - |

Table 4.6 – Coefficients de la combinaison linéaire des variables servant à calculer la première composante principale (de référence et simplifiée). Les $p^\star = 20$ premières variables sont rangées par ordre d'importance dans les meilleurs sous-ensembles.

Pour finir, et de manière indépendante, nous avons cherché à obtenir des indicateurs composites plus simples à écrire, c'est à dire écrits comme une combinaison linéaire d'un nombre restreint de variables. La méthode CSS a permis d'obtenir des indicateurs composites simplifiés et qui sont fortement corrélés aux indicateurs composites de référence.
Une autre approche méthodologique aurait pu être de réaliser la typologie des communes sur les indicateurs simplifiés, cependant nous avons choisi de réaliser la typologie sur les indicateurs de référence afin de conserver le plus d'information possible. Néanmoins, la typologie sur les indicateurs simplifiés a été effectuée mais n'est pas représentée ici. Cette typologie obtenue est naturellement très ressemblante à la typologie obtenue avec les indicateurs composites de référence.

# 5 Classification avec contraintes géographiques : la méthode hclustgeo

## Sommaire



## 5.1 Introduction

Nous avons vu dans les Chapitres 3 et 4 deux méthodes permettant de construire des indicateurs composites. Pour chacune de ces deux méthodes, une fois les indicateurs construits, nous avons réalisé une typologie des communes étudiées sur ces indicateurs. Plusieurs méthodes de classifications d'observations existent, les plus connues étant la classification ascendante hiérarchique (CAH) et la méthode des k-means. Leur but est de rassembler dans une même classe les observations qui se ressemblent (du pointde vue des variables qui les décrivent) et que les observations les plus différentes se retrouvent dans des classes différentes. Lorsque les observations à classer sont des unités spatiales, comme dans notre cas où des communes sont décrites par des indicateurs, il arrive

souvent que la typologie (partition) observée soit très fragmentée géographiquement. Or, il semble judicieux de penser que deux communes contiguës subissent a priori des influences semblables et auront donc tendance à se ressembler. De plus, la compréhension d'une typologie plus compacte d'un point de vue géographique peut être facilitée. Pour diminuer ce phénomène de fragmentation géographique et ainsi faire en sorte que des communes proches géographiquement aient plus de chances de se retrouver dans une même classe, il est nécessaire d'inclure de l'information spatiale dans la procédure de classification. Dans la littérature, il existe différentes solutions afin d'intégrer de l'information spatiale. Une revue de certaines de ces méthodes peut être trouvée dans Murtagh (1985) et Gordon (1996).

Il existe deux approches principales pour introduire de l'information spatiale dans la classification. La première approche consiste à utiliser une matrice binaire de contiguïté $\mathbf{Q}$ qui contient les relations de voisinage entre les observations ($\mathbf{Q}_{ij} = 1$ si les observations $x_i$ et $x_j$ sont voisines et 0 sinon). Cela requiert de définir la notion de voisinage entre observations : est ce que les observations $x_i$ et $x_j$ sont voisines si elles partagent une frontière commune ? Sont elles voisines si elles se trouvent à une certaine distance l'une de l'autre ? Il existe un grand nombre de manières, plus ou moins complexes, pour définir une relation de voisinage.

Plusieurs méthodes de classification utilisent une matrice $\mathbf{Q}$ de voisinage. Legendre and Legendre (2012) ont développé une méthode de CAH qui autorise uniquement l'agrégation d'observations voisines. La méthode proposée par Guo (2009) commence par effectuer une CAH avec contraintes basées sur la matrice de voisinage. Puis, cette partition est optimisée en changeant certaines observations de classes jusqu'à maximisation d'un certain critère d'homogénéité. Chavent et al. (2008) ont développé une méthode divisive de classification hiérarchique qui optimise un critère basé sur la distance entre les observations calculée à partir des variables mais également sur la matrice de voisinage $\mathbf{Q}$. Une autre méthode proposée par Ambroise et al. (1997) est une approche de type modèle basée également sur la matrice de voisinage $\mathbf{Q}$. Ils utilisent une modification de l'algorithme EM pour rassembler les observations dans des classes homogènes et géographiquement compactes.

La seconde approche la plus connue pour introduire de l'information spatiale dans la classification d'observations consiste à utiliser une matrice symétrique de distances géographiques entre les observations. Webster (1977) et Oliver and Webster (1988) définissent une mesure d'agrégation entre observations incluant une fonction non linéaire des distances géographiques entre observations. Cette mesure d'agrégation modifiée favorise le rassemblement de deux observations géographiquement proches et pénalise l'agrégation d'observations éloignées géographiquement. A notre connaissance, aucune des méthodes présentées précédemment ne sont implémentées dans un package R.

Ce chapitre présente la méthode hclustgeo, une nouvelle méthode de CAH incluant de l'information spatiale. Cette méthode est incluse dans un nouveau package R appelé `ClustGeo`. Les contraintes spatiales sont introduites par le biais d'une matrice de distance géographiques entre observations appelée $\mathbf{D}_2$. Dans toute la suite, on considère que la matrice $\mathbf{D}_2$ est une matrice de distances euclidiennes. Parallèlement à cette matrice $\mathbf{D}_2$, on utilise également la matrice $\mathbf{X}$ de dimension $n \times p$ où $n$ observations $x_1, \ldots, x_n$ sont décrites par $p$ variables quantitatives. A partir de cette matrice $\mathbf{X}$, on calcule la matrice de distances euclidiennes entre les observations que l'on note $\mathbf{D}_1$. La méthode hclustgeo peut être vue comme un compromis entre une CAH de Ward effectuée uniquement sur la matrice $\mathbf{D}_1$ et une CAH de Ward effectuée uniquement sur la matrice $\mathbf{D}_2$. Le but de la méthode est d'obtenir des classes d'observations homogènes (du point de vue des $p$ variables), mais qui soient plus compactes géographiquement. Un des avantages du package `ClustGeo` est qu'il peut créer des cartes géographiques de différentes typologies d'observations. Pour cela, il est nécessaire d'avoir les shapefiles associés aux observations. Un shapefile, est un format de fichier issu du monde des systèmes d'informations géographiques (SIG) permettant de créer et d'afficher des cartes. L'affichage des cartes est réalisé grâce à la fonction `plot.hclustgeo` du package. Cet affichage est possible grâce au package `rCarto`, voir Giraud (2013).

Ce chapitre est organisé de la manière suivante. La Section 5.2 explique le principe de la CAH avec critère additif d'hétérogénéité puis présente le cas de la CAH de Ward. La Section 5.3 présente la méthode hclustgeo et l'algorithme utilisé. La Section 5.4 est dédiée à l'illustration et l'utilisation des principales fonctions du package `ClustGeo` sur un exemple simple. La Section 5.5 concerne l'application de la méthode hclustgeo sur la zone d'étude des SAGEs présentée à la Section 4.4.

## 5.2 CAH avec critère additif d'hétérogénéité

Nous allons présenter dans cette section comment réaliser de manière générale une CAH lorsque le critère choisi pour mesurer l'hétérogénéité d'une partition est additif. Puis, dans un second temps nous présenterons le cas particulier de la CAH avec critère additif de Ward, ce critère étant au coeur de la méthode hclustgeo.

### 5.2.1 Présentation générale

Nous présentons ici la méthodologie générale pour construire une hiérarchie indicée lorsque le critère d'hétérogénéité d'une partition est additif.

**Critère d'hétérogénéité additif.** On note $\mathcal{H}(\mathcal{P}_K)$ le critère additif d'hétérogénéité de la partition $\mathcal{P}_K$ en $K$ classes. Ce critère est défini comme la somme des hétérogénéités des classes $H(\mathcal{C}_k)$ :

$$\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} H(\mathcal{C}_k). \tag{5.2.1}$$

**Algorithme.** L'algorithme de CAH débute avec la partition en singletons. A chaque étape de l'algorithme, on sélectionne deux classes à agréger. Cela conduit à l'étape d'après à une partition contenant une classe en moins. L'algorithme est terminé lorsque la partition en une seule classe, notée $\mathcal{P}_1$, contenant toutes les observations est obtenue. A chaque étape $s = 1 \dots n-2$, on agrège les deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ de la partition $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ telles que leur rassemblement produise la partition $\mathcal{P}_{K-1}$ en $K-1$ classes ayant le plus petit critère d'hétérogénéité $\mathcal{H}(\mathcal{P}_{K-1})$. Cette étape d'agrégation est ainsi basée sur une mesure d'agrégation $\delta(\mathcal{C}_l, \mathcal{C}_m)$ entre deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$, définie comme la différence d'hétérogénéité entre la partition en $K$ classes et la partition en $K-1$ classes (issue du rassemblement des classes $\mathcal{C}_l$ et $\mathcal{C}_m$) :

$$\delta(\mathcal{C}_l, \mathcal{C}_m) = \mathcal{H}(\mathcal{P}_{K-1}) - \mathcal{H}(\mathcal{P}_K) = H(\mathcal{C}_l \cup \mathcal{C}_m) - H(\mathcal{C}_l) - H(\mathcal{C}_m). \tag{5.2.2}$$

A chaque étape de l'algorithme, on rassemble les deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ telles que $\delta(\mathcal{C}_l, \mathcal{C}_m) = \min_{i,j} \delta(\mathcal{C}_i, \mathcal{C}_j)$.

En sortie de l'algorithme, on obtient une hiérarchie que l'on peut indicer par $h$ avec $h(\mathcal{C}_l \cup \mathcal{C}_m) = \delta(\mathcal{C}_l, \mathcal{C}_m)$. Cette hiérarchie peut alors être représentée par un dendrogramme. Une suite de partitions emboîtées de 2 à $n$ classes, peut alors être obtenue en coupant le dendrogramme suivant une suite de lignes horizontales.

**Critère de qualité d'une partition.** Un critère de qualité d'une partition $\mathcal{P}_K$ en $K$ classes peut alors être défini comme suit :

$$1 - \frac{\mathcal{H}(\mathcal{P}_K)}{\mathcal{H}(\mathcal{P}_1)}, \tag{5.2.3}$$

où $\mathcal{P}_1$ est la partition avec toutes les observations dans une seule classe. Ce critère varie entre zéro et un. Il est égal à 1 pour la partition en $n$ classes des singletons et il vaut zéro pour la partition $\mathcal{P}_1$ en une seule classe. Ce critère augmente avec le nombre de classes, ainsi il n'est utile que pour comparer des partitions ayant le même nombre de classes. L'utilisateur choisira la partition avec le plus grand critère de qualité.

### 5.2.2 Exemple de la CAH avec critère de Ward

La CAH de WARD est un exemple bien connu de CAH avec critère additif d'hétérogénéité. Dans ce cas, le critère d'hétérogénéité d'une classe $\mathcal{C}_k$ est défini comme l'inertie (notée $I$) des observations appartenant à la classe. On se place dans le cas où toutes les observations ont un poids identique égal à $\frac{1}{n}$. L'hétérogénéité d'un cluster $\mathcal{C}_k$ est donnée par :

$$H(\mathcal{C}_k) = I(\mathcal{C}_k) = \sum_{i \in \mathcal{C}_k} \frac{1}{n} d^2(x_i, g_k), \tag{5.2.4}$$

où $d^2(x_i, g_k)$ est le carré de la distance euclidienne entre l'observation $x_i$ et le centre de gravité $g_k = \sum_{x_i \in \mathcal{C}_k} \frac{1}{n} x_i$ de la classe $\mathcal{C}_k$.
On peut également calculer l'inertie des observations d'une classe sans calculer les coordonnées de son centre de gravité mais uniquement en utilisant l'ensemble des distances prises deux à deux entre les observations de la manière suivante :

$$H(\mathcal{C}_k) = I(\mathcal{C}_k) = \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \frac{1}{2nn_k} d^2(x_i, x_j), \tag{5.2.5}$$

où $n_k$ est le nombre d'observations appartenant à la classe $\mathcal{C}_k$.

L'hetérogénéité $\mathcal{H}(\mathcal{P}_K)$ de la partition $\mathcal{P}_K$ est alors égale à l'inertie intra-classe de la partition $W(\mathcal{P}_K)$, où $W(\mathcal{P}_K) = \sum_{k=1}^{K} I(\mathcal{C}_k)$. Avec ce critère d'hétérogénéité, la mesure d'agrégation (appelée mesure d'agrégation de Ward) entre deux classes peut être écrite de la manière suivante :

$$\delta_{Ward}(\mathcal{C}_l, \mathcal{C}_m) = I(\mathcal{C}_l \cup \mathcal{C}_m) - I(\mathcal{C}_l) - I(\mathcal{C}_m) = \frac{1}{n} \frac{n_l \times n_m}{n_l + n_m} d^2(g_l, g_m), \tag{5.2.6}$$

avec $n_l$ (resp. $n_m$) le nombre d'observations dans la classe $\mathcal{C}_l$ (resp. $\mathcal{C}_m$), $g_l$ (resp. $g_m$) le centre de gravité de la classe $\mathcal{C}_l$ (resp. $\mathcal{C}_m$) et $d^2(g_l, g_m)$ le carré de la distance euclidienne entre $g_l$ et $g_m$.

**Calcul des mesures d'agrégation.** Dans l'algorithme de CAH, une fois que deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ ont été agrégées, il est nécessaire de calculer les mesures d'agrégation entre la nouvelle classe formée $(\mathcal{C}_l \cup \mathcal{C}_m)$ et toutes les autres classes. Nous avons vu à l'équation (5.2.6) que pour calculer la mesure d'agrégation de Ward entre deux classes, il est nécessaire de calculer leurs centres de gravité respectifs, ceci nécessite de revenir à la matrice de données et peut être couteux en temps de calcul. En pratique, on peut éviter de calculer les coordonnées des centres de gravité à chaque étape. En effet, il est possible de calculer la mesure d'agrégation entre la nouvelle classe $\mathcal{C}_l \cup \mathcal{C}_m$ et toutes les

autres classes $\mathcal{C}_k$ uniquement à l'aide des mesures d'agrégation entre les trois classes $\mathcal{C}_l$, $\mathcal{C}_m$ et $\mathcal{C}_k$. Ceci est possible en utilisant la formule de *Lance et Williams*, voir Lance and Williams (1967), qui s'écrit de la manière suivante, dans le cas de la mesure de Ward :

$$\delta_{Ward}(\mathcal{C}_l \cup \mathcal{C}_m, \mathcal{C}_k) = \frac{n_l + n_k}{n_l + n_m + n_k}\delta_{Ward}(\mathcal{C}_l, \mathcal{C}_k) + \frac{n_m + n_k}{n_l + n_m + n_k}\delta_{Ward}(\mathcal{C}_m, \mathcal{C}_k) - \frac{n_k}{n_l + n_m + n_k}\delta_{Ward}(\mathcal{C}_l, \mathcal{C}_m). \quad (5.2.7)$$

L'utilisation de cette formule permet d'accélérer considérablement l'algorithme. En effet, les centres de gravité des classes n'ont plus besoin d'être calculés à chaque itération, seules les mesures d'agrégation entre les classes à l'étape $s$ sont nécessaires pour obtenir les mesures d'agrégation entre les classes à l'étape $s+1$. Ainsi, nous avons juste besoin de calculer toutes les mesures d'agrégation entre les paires d'observations à la première étape puis de les mettre à jour grâce à l'équation (5.2.7). La mesure d'agrégation de Ward entre deux singletons $\{x_i\}$ et $\{x_j\}$ est donnée par :

$$\delta_{Ward}(\{x_i\}, \{x_j\}) = \frac{1}{2n}d^2(x_i, x_j). \quad (5.2.8)$$

On calcule donc toutes les mesures d'agrégation entre singletons dans la matrice $\mathbf{\Delta} = \frac{1}{2n}\mathbf{D}_1^2$, où $\mathbf{D}_1$ est la matrice $n \times n$ contenant les distances euclidiennes entre toutes les observations (on rappelle que $\mathbf{D}_1$ est calculée à partir de la matrice de données $\mathbf{X}$). Une fois la matrice $\mathbf{\Delta}$ calculée, on la met à jour à chaque étape de l'algorithme grâce à l'équation (5.2.7). On remarque que tout l'algorithme de la CAH de Ward peut être mis en oeuvre uniquement en calculant la matrice $\mathbf{D}_1$ de distances euclidiennes entre les observations à partir de la matrice $\mathbf{X}$.

**Qualité d'une partition obtenue avec la mesure d'agrégation de Ward.** A chaque étape d'agrégation de l'algorithme de CAH avec critère de Ward, on minimise l'augmentation de l'inertie intra-classe. Le critère de qualité de partition défini à l'équation (5.2.3) est donc égal à $1 - \frac{W(\mathcal{P}_K)}{I(\mathcal{P}_1)}$, le pourcentage d'inertie expliquée par la partition $\mathcal{P}_k$.

## 5.3 La méthode hclustgeo

On cherche ici à intégrer l'information géographique sur les observations contenue dans la matrice de distances géographiques $\mathbf{D}_2$. Pour cela, nous définissons une nouvelle mesure d'hétérogénéité de classe puis nous appliquons un algorithme de CAH avec critère additif. Cela amène également la définition d'une nouvelle mesure d'agrégation $\delta(\mathcal{C}_l, \mathcal{C}_m)$ entre classes.

**Hétérogénéité d'une classe.** Soit $\alpha \in [0,1]$. On considère ici :

$$H(\mathcal{C}_k) = \alpha I(\mathcal{C}_k, \mathbf{D}_1) + (1-\alpha) I(\mathcal{C}_k, \mathbf{D}_2), \tag{5.3.1}$$

où $I(\mathcal{C}_k, \mathbf{D}_1)$ (resp. $I(\mathcal{C}_k, \mathbf{D}_2)$) est l'inertie des observations de la classe $\mathcal{C}_k$ calculée à partir de la matrice de distances euclidiennes $\mathbf{D}_1$ (resp. $\mathbf{D}_2$) comme vue à l'équation (5.2.5). Avant de procéder à ces calculs d'inertie à l'aide des matrices $\mathbf{D}_1$ et $\mathbf{D}_2$, il est nécessaire que celles-ci aient le même ordre de grandeur. Pour cela, nous avons choisi de les normaliser de la manière suivante : $\mathbf{D}_1 \leftarrow \frac{\mathbf{D}_1}{max(\mathbf{D}_1)}$ et $\mathbf{D}_2 \leftarrow \frac{\mathbf{D}_2}{max(\mathbf{D}_2)}$. Cependant, d'autres types de normalisation auraient pu être envisagés. Une fois la normalisation effectuée, le critère (5.3.1) permet de donner plus ou moins d'importance à la distance géographique ou à l'information apportée par les variables, en fonction de la valeur du paramètre $\alpha$.

**Mesure d'agrégation $\delta$ entre deux classes.** Le critère d'hétérogénéité défini précédemment étant additif, on définit la mesure d'agrégation entre deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ de la même manière qu'à l'équation (5.2.2). On obtient ainsi la mesure d'agrégation suivante :

$$\begin{aligned}
\delta_{hgeo}(\mathcal{C}_l, \mathcal{C}_m) &= H(\mathcal{C}_l \cup \mathcal{C}_m) - H(\mathcal{C}_l) - H(\mathcal{C}_m) \\
&= \alpha \left[ I(\mathcal{C}_l \cup \mathcal{C}_m, \mathbf{D}_1) - I(\mathcal{C}_l, \mathbf{D}_1) - I(\mathcal{C}_m, \mathbf{D}_1) \right] \\
&\quad + (1-\alpha) \left[ I(\mathcal{C}_l \cup \mathcal{C}_m, \mathbf{D}_2) - I(\mathcal{C}_l, \mathbf{D}_2) - I(\mathcal{C}_m, \mathbf{D}_2) \right] \\
&= \alpha \ \delta^1_{Ward}(\mathcal{C}_l, \mathcal{C}_m) + (1-\alpha) \ \delta^2_{Ward}(\mathcal{C}_l, \mathcal{C}_m).
\end{aligned} \tag{5.3.2}$$

La mesure d'agrégation définie ici correspond à deux mesures d'agrégation de Ward (voir équation (5.2.6)) calculées à l'aide de deux matrices de distances différentes ($\mathbf{D}_1$ et $\mathbf{D}_2$) et pondérées respectivement par $\alpha$ et $1-\alpha$.
Ainsi, quand $\alpha = 1$, la méthode hclustgeo est équivalente à la CAH de Ward usuelle effectuée à l'aide de la matrice de distance euclidienne $\mathbf{D}_1$ calculée sur les $p$ variables. Inversement, quand $\alpha = 0$, la méthode hclustgeo est équivalente à la CAH de Ward effectuée sur la matrice de distances géographiques $\mathbf{D}_2$.

**Calcul des mesures d'agrégation.** Nous avons vu que la mesure d'agrégation $\delta_{hgeo}$ est une somme pondérée de mesures de Ward calculées sur des matrices de distances différentes. Ainsi, à chaque étape de l'algorithme, pour mettre à jour les mesures d'agrégation $\delta_{hgeo}$ entre la classe nouvellement créée et les autres classes, nous pouvons utiliser l'équation (5.2.7) pour mettre à jour séparément les mesures $\delta^1_{Ward}$ et $\delta^2_{Ward}$.
Ainsi, après avoir calculé les matrices $\mathbf{\Delta}_1 = \frac{1}{2n}{\mathbf{D}_1}^2$ et $\mathbf{\Delta}_2 = \frac{1}{2n}{\mathbf{D}_2}^2$ contenant les mesures de Ward calculées entre toutes les observations à l'aide des matrices de distance $\mathbf{D}_1$ et $\mathbf{D}_2$, on construit la matrice $\mathbf{\Delta} = \alpha\mathbf{\Delta}_1 + (1-\alpha)\mathbf{\Delta}_2$. Puis à chaque étape de

l'algorithme, après agrégation de deux classes, on met à jour séparément les matrices $\mathbf{\Delta}_1$ et $\mathbf{\Delta}_2$ à l'aide de l'équation (5.2.7) puis on reconstruit la matrice $\mathbf{\Delta}$ comme expliqué précédemment. Ainsi, on remarque que les seules matrices dont on a besoin pour réaliser l'algorithme de hclustgeo sont les matrices $\mathbf{D}_1$ et $\mathbf{D}_2$.

**Choix du paramètre $\alpha$.** Le choix du paramètre $\alpha$ est très important en pratique. L'idée est de sélectionner une valeur de $\alpha$ de sorte que le pourcentage d'inertie expliquée calculé avec $\mathbf{D}_1$ ne soit pas trop dégradé. Lorsque $\alpha = 1$, une CAH de Ward basée uniquement sur $\mathbf{D}_1$ est réalisée et aucune information spatiale n'est prise en compte. Inversement, lorsque $\alpha = 0$, une CAH de Ward basée uniquement sur $\mathbf{D}_2$ est réalisée, cela mène à une partition très compacte du point de vue géographique (des observations proches géographiquement sont dans la même classe) mais sans aucune ressemblance (du point de vue de la matrice de données $\mathbf{X}$) entre les observations des mêmes classes. L'idée est d'obtenir une partition qui est la plus compacte possible d'un point de vue géographique mais sans perdre trop de pourcentage d'inertie expliquée calculé avec $\mathbf{D}_1$. En effet, ce pourcentage d'inertie expliquée, mesure la ressemblance des observations au sein d'une même classe du point de vue des variables de la matrice de données $\mathbf{X}$. Dans notre problématique, il est préférable d'avoir une partition potentiellement fragmentée d'un point de vue géographique (des observations proches géographiquement ne sont pas dans la même classe) mais avec des observations qui se ressemblent (sur $\mathbf{X}$) au sein d'une même classe (c'est le cas lorsque $\alpha = 1$) plutôt qu'une partition géographiquement très compacte mais avec aucune ressemblance entre les observations d'une même classe (c'est le cas lorsque $\alpha = 0$). Ainsi, pour choisir une valeur du paramètre $\alpha$ on cherchera à ne pas trop détériorer le pourcentage d'inertie expliquée calculé uniquement à l'aide de $\mathbf{D}_1$. Ce pourcentage est donc le critère de qualité utilisé pour le choix du paramètre $\alpha$, il est donné par :

$$1 - \frac{W_1(\mathcal{P}_K, \mathbf{D}_1)}{I(\mathcal{P}_1, \mathbf{D}_1)}, \tag{5.3.3}$$

où $W_1(\mathcal{P}_k, \mathbf{D}_1)$ est l'inertie intra-classe de $\mathcal{P}_K$ calculée avec $\mathbf{D}_1$ et $I(\mathcal{P}_1, \mathbf{D}_1)$ est l'inertie du nuage de toutes les observations aussi calculée avec la matrice de distances $\mathbf{D}_1$.
En pratique, on propose la procédure suivante pour déterminer la valeur du paramètre $\alpha$ à utiliser :

- Premièrement, il est nécessaire de choisir le nombre $K$ de classes de la partition. On peut s'aider pour cela du dendrogramme de la CAH de Ward effectuée uniquement sur la matrice de données $\mathbf{X}$ (cas où $\alpha = 1$).
- Une fois le nombre $K$ de classes choisi, on sélectionne la plus petite valeur du paramètre $\alpha$ de telle sorte que la qualité de la partition, définie à l'équation (5.3.3), soit la moins dégradée possible. Pour cela, on peut se fixer par exemple, un seuil

de 10% de perte de qualité à ne pas dépasser. Dans ce cas, la valeur optimale de $\alpha$ est définie comme suit :

$$\begin{aligned} &\alpha^\star := \text{argmin}_\alpha \, q_\alpha, \\ &s.c \; q_\alpha \geq 0.9 q_1 \end{aligned} \tag{5.3.4}$$

où $q_\alpha$ (resp. $q_1$) correspond à la qualité de la partition $\mathcal{P}_K$ obtenue avec $\alpha$ (resp. $\alpha = 1$).
On note également que l'inspection visuelle des cartes géographique des partitions obtenues avec différentes valeurs de $\alpha$ peut également être informative pour choisir la valeur de $\alpha$.

**Algorithme de classification de hclustgeo.** L'algorithme de hclustgeo consiste à construire une hiérarchie indicée définissant un ensemble de $n$ partitions emboitées. L'algorithme prend en entrée les trois arguments suivants : la matrice $\mathbf{X}$ des données de dimension $(n \times p)$ contenant la description de $n$ observations par $p$ variables quantitatives ; la matrice $\mathbf{D}_2$ de dimension $(n \times n)$ contenant les distances géographiques entre les observations et le paramètre $\alpha$. On suppose également que la matrice $\mathbf{D}_2$ est une matrice de distances euclidiennes. L'algorithme fonctionne de la manière suivante :

1. Etape $s = 0$ : initialisation.
   (a) Calculer la matrice $\mathbf{D}_1$ de dimension $(n \times n)$ contenant les distances euclidiennes entre les observations à partir de la matrice de données $\mathbf{X}$.
   (b) Afin d'avoir le même ordre de grandeur entre les matrices $\mathbf{D}_1$ et $\mathbf{D}_2$, elles sont normalisées de la manière suivante : $\mathbf{D}_1 \leftarrow \frac{\mathbf{D}_1}{max(\mathbf{D}_1)}$ et $\mathbf{D}_2 \leftarrow \frac{\mathbf{D}_2}{max(\mathbf{D}_2)}$.
   (c) Calculer les matrices $\mathbf{\Delta}_1 = \frac{1}{2n}\mathbf{D}_1^2$ et $\mathbf{\Delta}_2 = \frac{1}{2n}\mathbf{D}_2^2$.
   (d) Construire la matrice contenant les mesures d'agrégation entre les observations : $\mathbf{\Delta} = \alpha\mathbf{\Delta}_1 + (1-\alpha)\mathbf{\Delta}_2$.
2. Etape $s = 1, \ldots, n-2$ :
   (a) On agrège les deux classes de la partition en $n-s+1$ classes afin d'obtenir une partition en $n-s$ classes. Pour cela, on choisi d'agréger les classes $\mathcal{C}_l$ et $\mathcal{C}_m$ avec la plus petite mesure d'agrégation contenue dans $\mathbf{\Delta}$.
   (b) Mettre à jour la matrice $\mathbf{\Delta}$ en calculant la mesure d'agrégation entre la nouvelle classe $\mathcal{C}_l \cup \mathcal{C}_m$ et toutes les autres classes. Pour cela on met à jour séparément les deux matrices $\mathbf{\Delta}_1$ et $\mathbf{\Delta}_2$ à l'aide de l'équation (5.2.7) puis, on calcule la nouvelle matrice $\mathbf{\Delta} = \alpha\mathbf{\Delta}_1 + (1-\alpha)\mathbf{\Delta}_2$.
3. Etape $s = n-1$ : stop. La partition $\mathcal{P}_1$ en une seule classe est atteinte.

## 5.4 Illustration de hclustgeo à l'aide du package ClustGeo

Tout d'abord, cette section présente les principales fonctions du package ClustGeo. Par la suite, nous verrons comment utiliser ces fonctions à l'aide d'une illustration sur un exemple simple et reproductible car les données nécessaires sont incluses dans le package.

### 5.4.1 Les principales fonctions du package ClustGeo

Le package ClustGeo contient trois fonctions principales qui sont présentées ci-dessous :

- La fonction hclustgeo permet de réaliser la méthode hclustgeo basée sur la matrice de données $\mathbf{X}$ (argument data), la matrice $\mathbf{D}_2$ de distances géographiques entre les observations (argument D.geo) et une grille de paramètres $\alpha$ (argument alpha). Le résultat de la fonction est un objet de classe "hclustgeo", cet objet présente peu d'intérêt en tant que tel. La plupart des résultats numériques et graphiques seront obtenus en utilisant les fonctions plot.hclustgeo et summary.hclustgeo sur un objet de classe "hclustgeo".
- La fonction plot.hclustgeo utilise en entrée un objet de classe "hclustgeo". La fonction plot permet d'afficher trois types de graphiques différents, grâce à la valeur du paramètre choice :
  - Si choice=dendro, la fonction renvoie plusieurs dendrogrammes représentant chacun les hiérarchies obtenues pour les différentes valeurs de alpha.
  - Si choice=maps, la fonction renvoie les cartes géographiques des partitions en $K$ classes pour les différentes valeurs de alpha. Le nombre de classes choisi est contenu dans le vecteur K.range qui est un paramètre de la fonction plot. L'affichage des cartes géographiques nécessite la disponibilité de shapefiles relatifs à la zone étudiée, pour afficher ces cartes, on doit indiquer dans le paramètre path.shp le chemin d'accès vers ces shapefiles. De plus, l'identifiant des observations dans les shapefiles est contenu dans le paramètre name.ind.shp.
  - Si choice=quality, la fonction renvoie un graphique contenant les qualités de partitions (voir équation (5.3.3)) pour différentes valeurs de alpha et différentes valeurs de $K$ contenues dans K.range. Ce graphique est très utile pour choisir une valeur de $\alpha$ optimale, comme défini à l'équation (5.3.4).
- La fonction summary.hclustgeo prend également en entrée un objet de classe "hclustgeo". Elle permet de caractériser une partition en $K$ classes obtenues pour une certaine valeur de $\alpha$ à l'aide d'une base de données entrée dans le paramètre data.desc. Ainsi pour une partition donnée, on retrouve les valeurs

moyennes par classe des variables contenues dans `data.desc`. Ceci permet de donner du sens à une partition.

### 5.4.2 Illustration du package `ClustGeo` sur un exemple simple

Nous allons illustrer ici l'utilisation des principales fonctions du package `ClustGeo` sur un jeu de données simple disponible dans le package. Ce jeu de données contient la matrice $\mathbf{X}$ (appelée ici `data.303`) contenant la description de $n = 303$ communes (les observations) par $p = 4$ variables quantitatives. La description des variables utilisées pour cette illustration est donnée à la Table 5.1. Le package contient également la matrice $\mathbf{D}_2$ (appelée ici `Dgeo.303`) contenant les distances géographiques entre les communes. Nous avons vu que l'affichage des cartes géographiques nécessite les shapefiles relatifs aux communes étudiées, ces shapefiles sont également disponibles dans le répertoire d'installation du package. L'extrait de code ci-dessous montre comment charger les données présentes dans le package que nous allons utiliser dans l'illustration :

```
#chargement du package
library("ClustGeo")

#chargement des donnees necessaires pour utiliser hclustgeo
data(comm303)
X  <- comm303$data.303 #matrice de donnees X contenant les 4 variables
D2 <- comm303$Dgeo.303 #matrice D2 contenant les distances geo

#declaration de l'adresse du repertoire contenant les shapefiles pour
   afficher les cartes
chemin.shp <- file.path(path.package("ClustGeo"), "shapes/comm303")

#identifiant des communes dans les shapefiles
id.obs <- "INSEE_COM"
```

| Variable | Description |
|---|---|
| `agri.land` | Pourcentage de la commune couvert par des terres agricoles |
| `employ.rate.city` | Taux d'emploi communal |
| `graduate.rate` | Pourcentage de la population de la commune ayant un diplôme de type BAC |
| `housing.appart` | Pourcentage d'appartements dans la commune (parmi l'ensemble des logements) |

TABLE 5.1 – Description des variables utilisées dans l'illustration de `ClustGeo`.

Toutes les étapes de la méthodologie utilisée pour réaliser la méthode hclustgeo seront expliquées et détaillées avec des extraits de code.

**Étape 1 : Réaliser `hclustgeo` pour $\alpha = 1$ dans le but de choisir un nombre $K$ de classes.** La première étape consiste à réaliser la méthode pour $\alpha = 1$ (c'est à dire effectuer une CAH de Ward classique basée sur $\mathbf{D}_1$) dans le but de choisir un nombre

de classe à l'aide du dendrogramme de la hiérarchie obtenue avec $\alpha = 1$. Le code ci-dessous montre comment réaliser la méthode hclustgeo et afficher le dendrogramme correspondant :

```
#On lance hclustgeo pour alpha=1
res.alpha1 <- hclustgeo(data=X, D.geo=D2, alpha=1)

#Affichage du dendrogramme
plot(res.alpha1, choice="dendro")
```

Le dendrogramme représentant la hiérarchie obtenue pour $\alpha = 1$ est représenté à la Figure 5.1. Au vu de ce dendrogramme, on décide de retenir la partition en $K = 5$ classes d'observations.

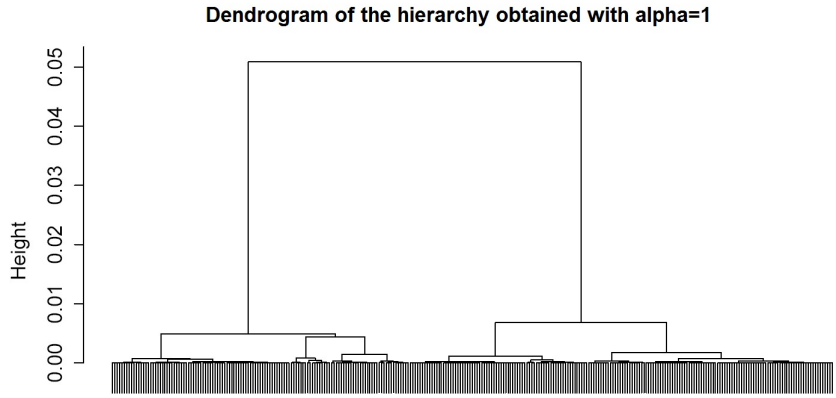



FIGURE 5.1 – Dendrogramme de la hiérarchie obtenue avec $\alpha = 1$.

**Étape 2 : Choix de la valeur du paramètre $\alpha$.** Dans le but de choisir une bonne valeur du paramètre $\alpha$, nous réalisons la méthode `hclustgeo` pour un ensemble de valeurs de $\alpha$. La "meilleure" valeur de $\alpha$ est choisie grâce à l'équation (5.3.4) qui donne le critère de qualité de partition défini comme le pourcentage d'inertie expliquée calculé uniquement à l'aide de la matrice $\mathbf{D}_1$. L'extrait de code R suivant détaille comment réaliser la méthode hclustgeo pour plusieurs valeurs de $\alpha$ puis comment obtenir les différentes qualités de partitions selon les valeurs de $\alpha$ et de $K$.

```
#on realise hclustgeo pour plusieurs valeurs de alpha
multi.alpha <- seq(0, 1, 0.1)
res.alpha <- hclustgeo(data=X, D.geo=D2, alpha=multi.alpha)

#graphique des qualites de partition pour K=3, 4 et 5 classes
plot.qual <- plot(res.alpha, choice="quality", K.range=c(3, 4, 5))
```

La Figure 5.2 représente les qualités de partitions en $K = 3, 4, 5$ obtenues avec différentes valeurs de $\alpha$. On note également que les valeurs numériques associées sont contenues dans l'objet `plot.qual`. Comme vu à l'étape 1, nous allons chercher la

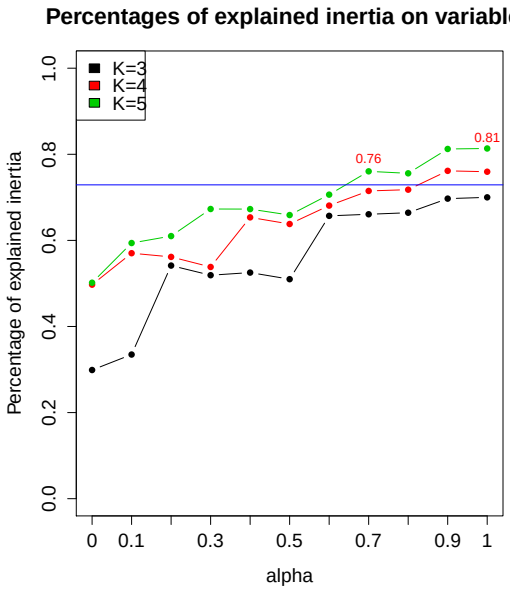


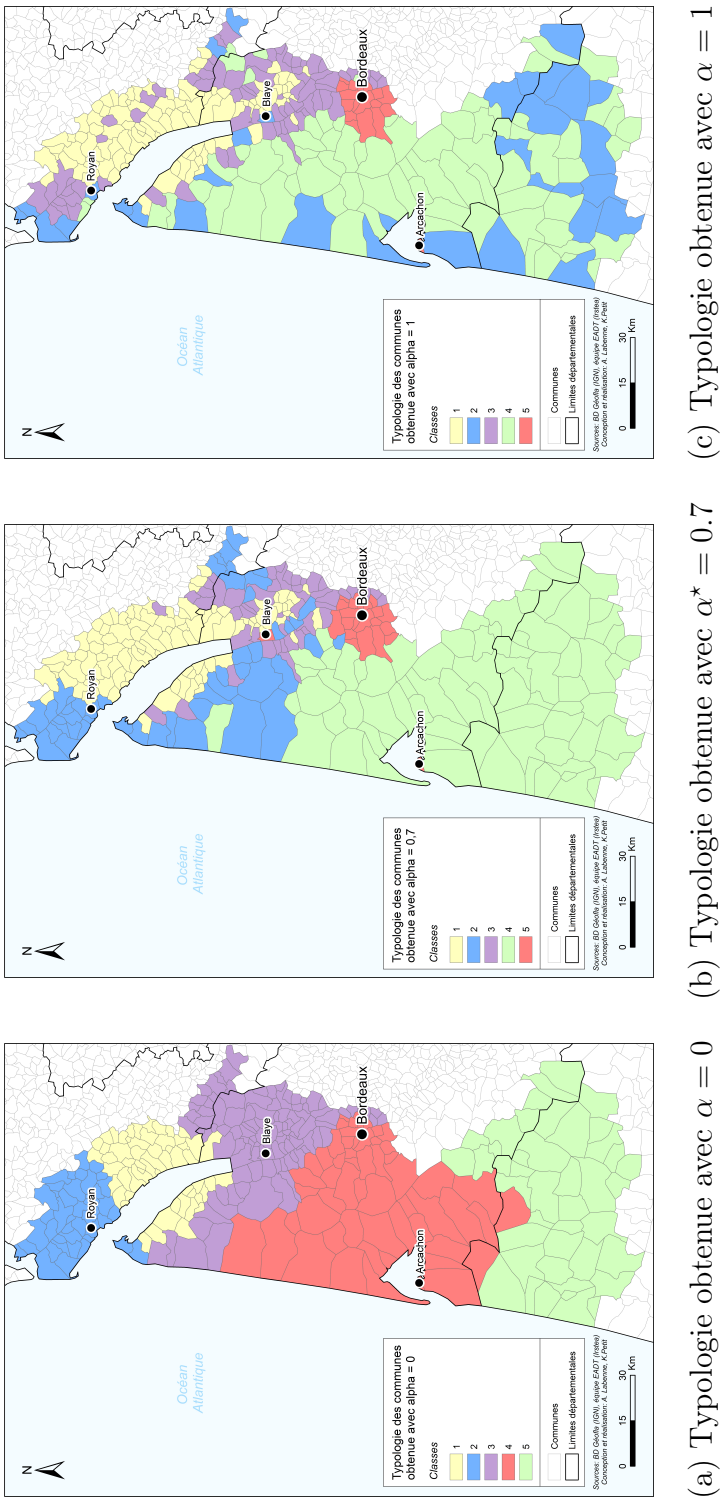

FIGURE 5.2 – Qualité des partitions en $K = 3, 4, 5$ classes en fonction des différentes valeurs de $\alpha$.

"meilleure" valeur $\alpha^\star$ pour une partition en $K = 5$ classes de communes. La ligne bleue sur le graphique correspond à une perte de qualité de 10% par rapport à la qualité de la partition en 5 classes obtenues avec $\alpha = 1$. On observe sur le graphique que pour $\alpha = 0.7$ on obtient une qualité de partition de 0.76 alors que la meilleure qualité de partition, obtenue avec $\alpha = 1$, est de 0.81. En choisissant la partition en 5 classes obtenues avec $\alpha^\star = 0.7$ on ne perd que 5% d'inertie expliquée calculée sur les variables par rapport à la partition obtenue avec $\alpha = 1$. On remarque également que si nous avions choisi les partitions en 3 ou 4 classes, nous aurions choisi la même valeur de $\alpha^\star$.

**Etape 3 : Affichage de la carte de la nouvelle typologie et comparaison des deux partitions.** Le code suivant permet d'afficher la carte de la partition en 5 classes obtenue avec $\alpha^\star = 0.7$ que nous comparons avec la carte de la partition obtenue avec $\alpha = 1$. Nous représentons également de manière illustrative la partition en 5 classe obtenue avec $\alpha = 0$, c'est à dire lorsque l'on réalise une CAH avec critère de Ward uniquement à l'aide de la matrice $\mathbf{D}_2$. Ces trois cartes sont représentées à la Figure 5.3, le code R suivant permet de les obtenir :

```
#affichage des cartes pour alpha= 0 ; 0.7 et 1
plot(res.alpha, choice="maps", K.range=5, choice.alpha=c(0, 0.7, 1),
     path.shp=chemin.shp, name.ind.shp=id.obs)
```

On observe sur la Figure 5.3 que la partition obtenue avec $\alpha = 0$ est très compacte d'un point de vue géographique, cependant elle n'a que peu d'intérêt car les communes dans une même classe ne se ressemblent pas du tout du point de vue de la matrice




(a) Typologie obtenue avec $\alpha = 0$ (b) Typologie obtenue avec $\alpha^\star = 0.7$ (c) Typologie obtenue avec $\alpha = 1$

FIGURE 5.3 – Carte des partitions obtenues avec $\alpha = 0$, $\alpha^\star = 0.7$ et $\alpha = 1$.

de données $\mathbf{X}$. Quand on observe la carte de la partition obtenue avec $\alpha^\star = 0.7$ on remarque que celle-ci est plus compacte géographiquement que la carte de la partition obtenue avec $\alpha = 1$ et donc plus simple à interpréter. De plus, comme nous l'avons vu, en choisissant la partition obtenue avec $\alpha^\star = 0.7$ nous avons perdu très peu de qualité de partition.

Comme nous l'avons fait dans les chapitres précédents, afin de caractériser les classes d'une partition il est intéressant de regarder les valeurs moyennes par classes des variables ayant servi à créer la partition. Ces valeurs sont obtenues à l'aide du code R suivant :

```
#caracterisation des classes des partitions
#obtenues avec alpha= 0. et alpha=1
summ <- summary(res.alpha, K.range=5, choice.alpha=c(0.7, 1),
   data.desc=X)

#caracterisation de la partition obtenue avec alpha=0.7
summ$"summary_hclustgeo.alpha=0.7"$desc$"K=5"

#caracterisation de la partition obtenue avec alpha=1
summ$"summary_hclustgeo.alpha=1"$desc$"K=5"
```

L'interprétation des partitions obtenues ne sera pas effectué ici car cet exemple n'est qu'illustratif et n'a servi qu'à présenter les fonctions du package. Une interprétation des partitions sera effectuée à la section suivante.

## 5.5 Application de la méthode sur la typologie des communes des SAGEs

Dans cette section, nous allons utiliser la méthode hclustgeo sur les indicateurs composites construits à la Section 4.4.2 à l'aide de la méthode MFAmix. Nous avions réalisé une typologie, par CAH de Ward, en $K = 6$ classes sur ces indicateurs composites à la Section 4.4.3. Nous allons chercher à introduire de l'information géographique dans la classification des communes.

Dans un premier temps, nous représentons à la Figure 5.4 le graphique des qualités de partitions. Ce graphique est utilisé pour choisir la "meilleure" valeur du paramètre $\alpha$. La ligne bleue horizontale sur le graphique correspond à une perte de qualité de 10% par rapport à la qualité de la partition obtenue avec $\alpha = 1$. Au vu de ce graphique, on décide de retenir la partition obtenue avec $\alpha^\star = 0.6$. En effet avec cette valeur de $\alpha^\star$, nous perdons peu de qualité de partition (0.57 pour $\alpha^\star = 0.6$ contre 0.62 pour la partition de référence obtenue avec $\alpha = 1$).

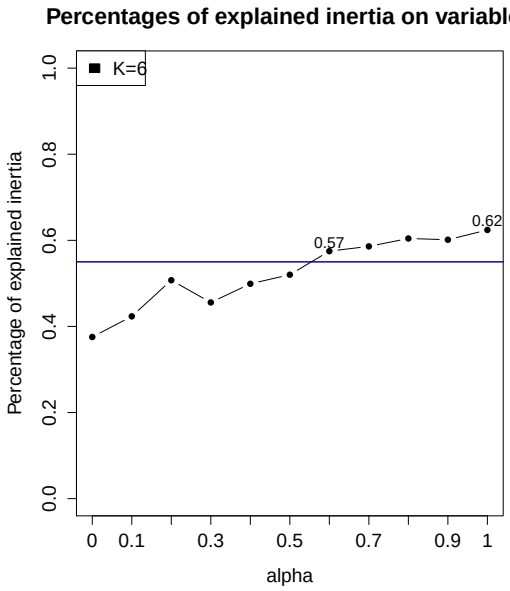



FIGURE 5.4 – Qualité de la partition en $K = 6$ classes en fonction des différentes valeurs de $\alpha$ sur la zone des SAGEs.

Nous avons représenté à la Figure 5.5 les cartes des partitions obtenues avec $\alpha = 1$ et $\alpha^\star = 0.6$. De plus, grâce à la fonction `summary` nous avons obtenu la caractérisation des classes de communes à l'aide des indicateurs composites de MFAmix. Cette caractérisation des classes est donnée à la Table 5.2.
L'examen de ces cartes montre tout d'abord que la partition obtenue avec $\alpha^\star = 0.6$ est plus compacte du point de vue géographique que la partition obtenue par CAH de Ward sur $\mathbf{D}_1$ (cas où $\alpha = 1$). L'examen des cartes et des valeurs moyennes des indicateurs par classe montre que la classe 5 représentant le pôle urbain de Bordeaux est exactement la même dans les deux partitions, cela montre une structure forte au sein de cette classe. On observe également que la classe 6 obtenue avec $\alpha^\star = 0.6$ située au tour du pôle urbain de Bordeaux est très ressemblante à la classe 6 obtenue avec CAH de Ward (cas où $\alpha = 1$). On observe la même chose avec la classe 5 qui se situe en majorité autour de Blaye sur les deux partitions. C'est aussi le cas pour la classe 4 qui est située en majorité au sud de la zone d'étude, ce qui correspond à la forêt des Landes. La classe 2 a légèrement changé sur la partition obtenue avec $\alpha^\star = 0.6$, cependant elle est toujours majoritairement située au nord de l'estuaire de la Gironde. La classe 1 a aussi changé mais sur les deux partitions, la plupart des communes de cette classe sont des communes littorales. De plus, lorsque l'on regarde les valeurs moyennes des indicateurs par classe, on remarque que les valeurs sont assez semblables pour les deux partitions. L'interprétation des classes de la typologie obtenue avec $\alpha = 1$ effectuée à

| **Indicateur composite** | **Classe de communes** | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| | *n=83* | *n=46* | *n=66* | *n=46* | *n=19* | *n=43* |
| CP 1 : Urbanisation | **-0.75** | **1.14** | **-1.48** | -0.31 | **4.70** | **0.76** |
| CP 2 : Emplois et ménages | **0.66** | **-1.91** | -0.44 | **-0.63** | 0.50 | **1.90** |
| CP 3 : Environnement | **0.88** | -0.43 | **0.34** | **-1.12** | **1.32** | **-1.15** |
| CP 4 : Qualification de l'emploi | **-0.29** | 0.19 | **0.84** | **-1.02** | 0.24 | 0.05 |

(a) Caractérisation de la partition obtenue avec $\alpha = 1$.

| **Indicateur composite** | **Classe de communes** | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| | *n=66* | *n=31* | *n=97* | *n=49* | *n=19* | *n=41* |
| CP 1 : Urbanisation | -0.27 | 0.28 | **-1.34** | -0.34 | **4.70** | **0.80** |
| CP 2 : Emplois et ménages | **0.72** | **-1.91** | -0.32 | **-0.95** | 0.50 | **1.97** |
| CP 3 : Environnement | **0.83** | -0.51 | **0.52** | **-1.34** | **1.32** | **-1.19** |
| CP 4 : Qualification de l'emploi | -0.33 | **0.71** | **0.29** | **-0.72** | 0.24 | 0.06 |

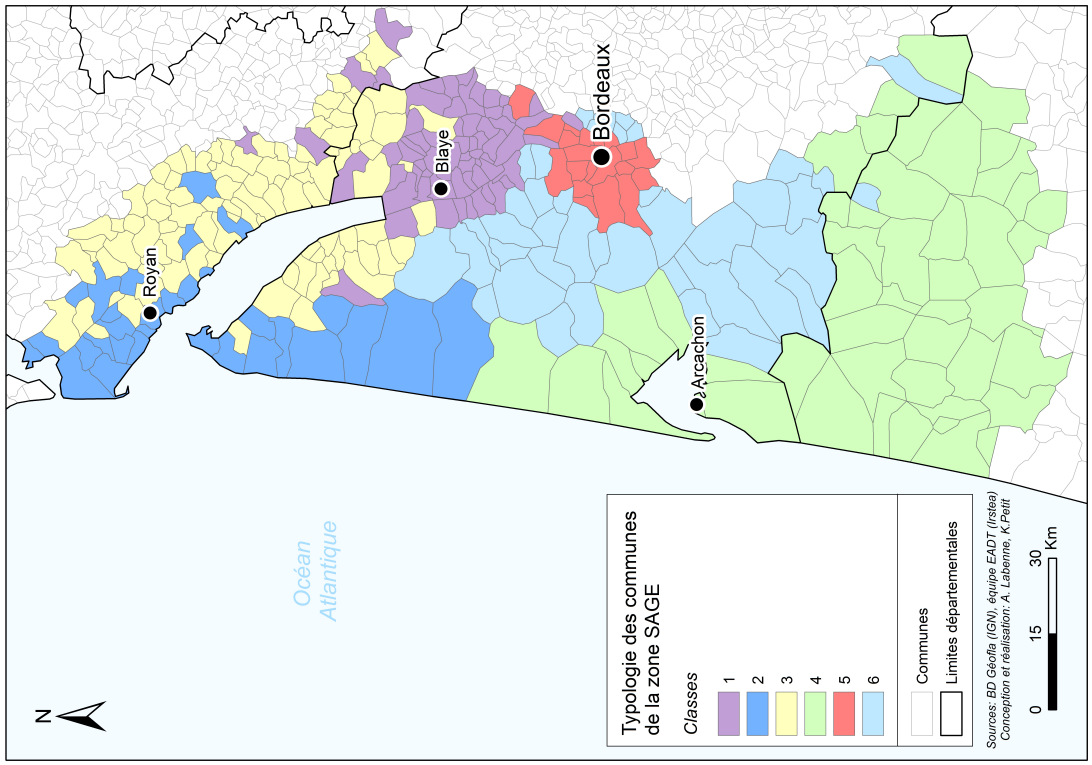
(b) Caractérisation de la partition obtenue avec $\alpha^\star = 0.6$.

TABLE 5.2 – Valeurs moyennes des indicateurs composites sur les cinq classes de communes obtenues avec $\alpha = 1$ et $\alpha^\star = 0.6$. En gras : valeurs significativement différentes de la moyenne globale de l'indicateur (p-value inférieure à $10^{-5}$). Par construction la moyenne globale est nulle.

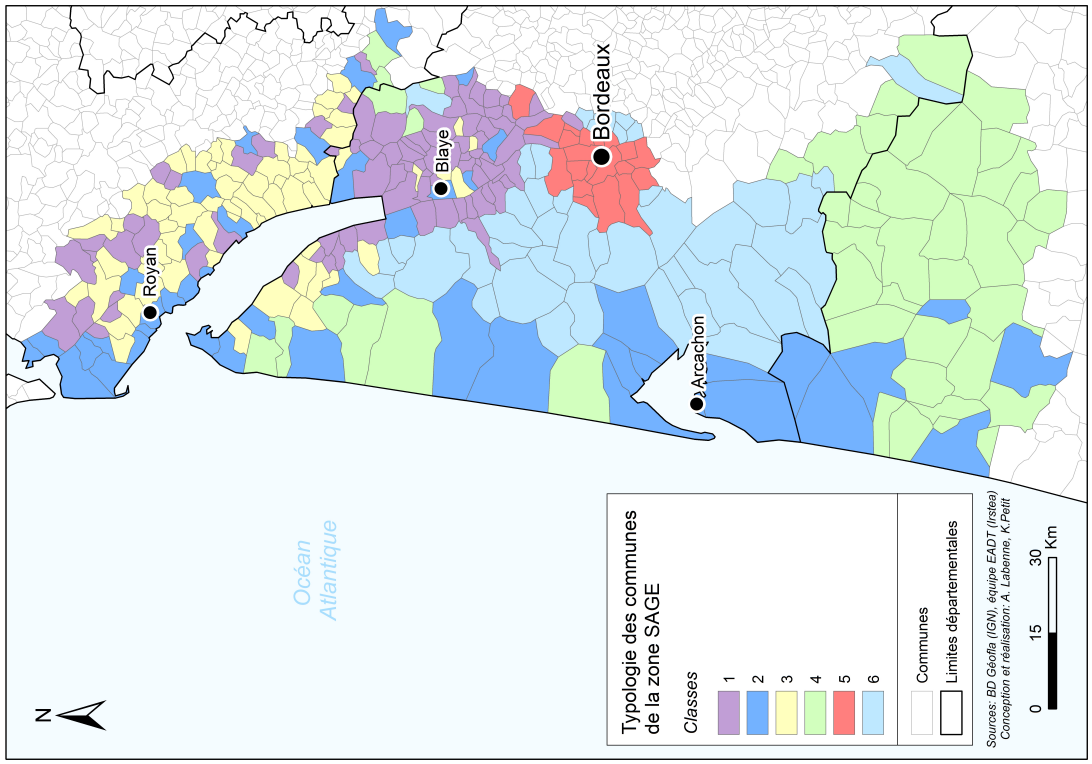
la Section 4.4.3 est donc très ressemblante à l'interprétation des classes de la typologie obtenue avec $\alpha^\star = 0.6$

## 5.6 Conclusion

Nous avons présenté dans ce chapitre une méthode de CAH permettant d'intégrer de l'information spatiale entre les observations. Cette méthode, appelée hclustgeo, est directement basée sur le critère de Ward et est implémentée dans le package `ClustGeo`. La méthode hclustgeo est très simple d'utilisation, en effet, elle nécessite une matrice de données $\mathbf{X}$ de dimension $(n \times p)$, une matrice $\mathbf{D}_2$ de dimension $(n \times n)$ contenant les distances géographiques euclidiennes entre les observations et un paramètre $\alpha \in [0, 1]$ qui donnera plus ou moins d'importance à la matrice de données $\mathbf{X}$ ou à la matrice de distances géographiques $\mathbf{D}_2$. Le package `ClustGeo` contient également d'autres fonctions utilisées pour obtenir des partitions issues des hiérarchies obtenues avec hclustgeo mais également pour représenter ces partitions sur une carte géographique. Nous avons proposé une méthodologie pratique afin de sélectionner la meilleure valeur du paramètre $\alpha$ à utiliser. Les différentes fonctions du package et leur utilisation sur un exemple simple ont été détaillées. Enfin, nous avons appliqué la méthode hclustgeo afin d'introduire de l'information géographique au sein d'une classification de communes effectuée




(b) Carte de la typologie obtenue avec $\alpha^\star = 0.6$




(a) Carte de la typologie obtenue avec $\alpha = 1$

FIGURE 5.5 – Carte des partitions obtenues avec $\alpha = 1$ et $\alpha^\star = 0.6$.

sur des indicateurs composites de qualité de vie. La nouvelle typologie obtenue s'est avérée très satisfaisante dans le sens où elle était plus compacte d'un point de vue géographique et sa qualité était très légèrement inférieure à la qualité de la partition obtenue avec la CAH de Ward usuelle (réalisée sur $\mathbf{D}_1$). Le package `ClustGeo` est disponible sur le CRAN mais également sur Github où l'on peut trouver une version en cours d'amélioration. En effet nous souhaitons par la suite améliorer le package afin que l'utilisateur puisse utiliser en entrée de la fonction `hclustgeo` une matrice $\mathbf{X}$ contenant des variables quantitatives et des variables qualitatives. De plus nous travaillerons à améliorer l'ergonomie générale du package.

# 6 Conclusion générale et perspectives

## Sommaire



## 6.1 Développements théoriques et packages R associés

Dans ce travail de thèse, nous avons détaillé deux méthodes statistiques de réduction de dimension pouvant être utilisées pour la construction d'indicateurs composites : la méthode de classification de variables du package `ClustOfVar` et la méthode d'analyse factorielle multiple MFAmix. Un des avantages de ces deux méthodes est qu'elles peuvent prendre en compte un mélange de variables quantitatives et de variables qualitatives. En effet, ces deux méthodes sont basées sur la méthode d'analyse factorielle de données mixtes PCAmix.

La méthode PCAmix, présentée au Chapitre 2, peut être vue comme un mélange de l'analyse en composantes principales (ACP) et de l'analyse des correspondances multiples (ACM), elle a été développée par Chavent et al. (2012). Dans ce travail de thèse, la méthode a été réécrite à l'aide d'une décomposition en valeurs singulières généralisée (GSVD) afin de pouvoir l'utiliser au sein de la méthode MFAmix. De plus, nous avons également amélioré significativement son utilisation dans R au sein de la fonction `PCAmix` du package `Pcamixdata`. En effet, dans la première version du package, certaines sorties numériques n'étaient pas disponibles, c'est le cas par exemple des contributions et des cosinus carrés. Ces sorties numériques ont été intégrées dans le package et nous avons également amélioré les sorties graphiques, avec la possibilité par

exemple de ne représenter que certaines observations ou variables en fonction de leurs qualités de représentation.

La méthode de classification de variables du package `ClustOfVar` a été présentée au Chapitre 3. Au sein de ce travail de thèse, nous avons utilisé cette méthode, mais nous avons également amélioré le package `ClustOfVar`. En effet, d'une part la vitesse de l'algorithme de la fonction `hclustvar` a été améliorée. D'autre part, nous avons intégré de nouvelles sorties numériques et graphiques permettant de faciliter l'interprétation des variables synthétiques associées aux différents clusters de variables. Les résultats obtenus au Chapitre 3 ont été utilisés dans le cadre du projet ANR ADAPT'EAU, ils ont également fait l'œuvre d'un article en cours de révision au sein de la revue *Social Indicators Research*.

La méthode MFAmix présentée au Chapitre 4 que nous avons développée est une extension de l'analyse factorielle multiple (AFM), voir Escofier and Pagès (1983) et Bécue-Bertaut and Pagès (2008). En effet, l'AFM permettait seulement l'analyse de groupes de variables de même nature. L'extension de cette méthode, appelée MFAmix, nous a permis d'intégrer la mixité à l'intérieur des groupes, c'est à dire d'analyser des sous-tableaux contenant des variables quantitatives et des variables qualitatives. Nous avons détaillé l'écriture théorique de cette méthode ainsi que l'écriture des sorties numériques associées. De plus, nous avons également présenté une approche de sélection de variables permettant de réaliser MFAmix sur un ensemble réduit de variables, tout en obtenant des composantes principales très proches des composantes principales calculées sur l'ensemble des variables d'origine. La méthode MFAmix a également été intégrée au package R `PCAmixdata` au sein de la fonction `MFAMix`. De plus un travail d'homogénéisation a été effectué afin que son utilisation soit semblable à celle de la fonction `PCAmix`. Au cours de cette thèse, la méthode MFAmix a également été mobilisé sur une problématique différente de la qualité de vie mais en lien avec le projet ADAPT'EAU. En effet, la méthode a été utilisée afin d'étudier la morphologie de la Garonne a différentes dates, cette étude a fait l'objet d'un article en révision, coécrit avec Philippe Valette et Mélodie David du laboratoire GEODE à Toulouse, dans la revue *Geomorpghology* et actuellement en révision.

Pour finir, après avoir construit des indicateurs composites, il nous a semblé naturel de réaliser des typologies (ou partitions) des communes à l'aide de ces indicateurs. Pour cela, nous avons utilisé la méthode de classification ascendante hiérarchique (CAH) de Ward. Cependant, la cartographie des typologies a révélé des classes de communes fragmentée d'un point de vue géographique. Afin d'obtenir des classes plus compactes d'un point de vue géographique, nous avons décidé d'intégrer de l'information spatiale dans la procédure de classification. La méthode `hclustgeo` présentée au Chapitre 5 est une méthode de CAH intégrant une matrice de distances géographiques entre les com-

munes. Cette méthode permet, par le biais d'un paramètre $\alpha \in [0, 1]$ de donner plus ou moins d'importance à la distance géographique ou aux indicateurs et ainsi favoriser l'agrégation de deux communes proches géographiquement. Cette méthode a été implémentée dans le package `ClustGeo`. De plus, ce package permet également l'affichage des typologies obtenues sur une carte géographique à l'aide de fichiers spécifiques. La méthode hclustgeo fera l'objet d'un article qui sera soumis prochainement.

## 6.2 Discussion sur les méthodes proposées

Bien que les deux méthodes que nous avons utilisé ont permis la construction d'indicateurs composites de qualité de vie, elles sont assez différentes dans la manière d'analyser les données.

La méthode MFAmix permet d'équilibrer l'importance des différents groupes de variables dans la construction des indicateurs composites. Cependant, cela nécessite de définir des groupes de variables en fonction de la thématique à laquelle appartiennent les différentes variables. Ce choix a priori n'est pas toujours trivial à effectuer et nécessite une discussion avec des experts du domaine. De plus, nous avons vu que l'ensemble des indicateurs construits à l'aide de cette méthode sont tous une combinaison linéaire de toutes variables. Ceci peut être une contrainte pour l'interprétation et l'écriture des indicateurs composites.

L'approche par classification de variables est différente. En effet, cette méthode prend en compte l'ensemble des variables sans tenir compte d'une éventuelle structure en groupes. Elle permet de rassembler dans des clusters homogènes les variables apportant la même information. De plus chaque cluster de variables est représenté par une variable synthétique qui est la plus liée aux variables du cluster. Cette variable synthétique peut donc être vue comme un indicateur composite relatif aux variables du cluster. Ainsi, l'un des avantages de l'approche par classification de variables est que les indicateurs composites sont plus simples à écrire car ils s'écrivent comme une combinaison linéaire des seules variables du cluster considéré. En effet, chacune des variables n'intervient que dans la construction d'un seul indicateur. Ceci facilite l'écriture et la compréhension des indicateurs composites.

Les deux approches développées dans cette thèse pour la construction d'indicateurs composites ont permis de mieux comprendre la structuration des différentes dimensions de la qualité de vie. Le fait d'intégrer un grand nombre de variables a permis de dépasser l'approche consistant à choisir a priori un faible nombre d'indicateurs simples, ainsi le caractère multidimensionnel de la qualité de vie a pu être pris en compte. Ces deux méthodes ont été utilisées sur deux zones d'étude différentes et ont permis d'effectuer un diagnostic socio-économique et environnemental dans le but d'appréhender

la vulnérabilité des territoires en question.

## 6.3 Perspectives futures

Une des premières perspectives d'évolution de ce travail concerne les méthodes de rotation en analyse factorielle. Les méthodes de rotation permettent de faciliter l'interprétation des composantes principales. Le principe de ces méthodes est d'appliquer une rotation sur les premières composantes retenues. Cette procédure permet d'obtenir de nouvelles composantes principales plus clairement reliées ou non aux variables et donc plus facilement interprétables. Cette méthode a été entièrement écrite pour la méthode PCAmix, voir Chavent et al. (2012), et est incluse dans le package `PCAmixdata` au sein de la fonction `PCArot`. Il serait intéressant d'étendre cette méthode de rotation aux composantes principales issues de MFAmix. Ainsi, nous pourrons obtenir des indicateurs composites plus simples à comprendre.

Une seconde perspective d'évolution concerne la réduction du nombre de variables dans MFAmix. Nous avons présenté la méthode CSS permettant d'obtenir des composantes principales calculées sur un ensemble restreint de variables. Les méthodes dites parcimonieuses, comme l'ACP "sparse" par exemple, permettent d'obtenir des composantes principales plus simples que l'ACP classique. En effet le but de l'ACP "sparse" est d'introduire des contraintes afin d'obtenir des composantes principales s'écrivant comme une combinaison linéaire d'un nombre très restreint de variables. Ces méthodes parcimonieuses pourrait être étendues au cas de la méthode MFAmix, toujours dans le but d'obtenir des indicateurs composites plus simple à interpréter.

Une autre perspective d'amélioration concerne la prise en compte de l'information spatiale entre les communes. En effet, nous avons essayé d'intégrer cette information dans la procédure de classification. Il pourrait être intéressant d'introduire cette information directement au sein des méthodes de construction des indicateurs composites.

# Liste des travaux

## Articles à paraître dans des revues internationales à comité de lecture

— Mélodie David, Amaury Labenne, Jean-Michel Carozza, Philippe Valette. Trajectory of change in the middle Garonne river during the last 150 years : learning of mixed Multiple Factor Analysis (MFAmix) in the study of historical maps. A paraître dans *Geomorphology*

— Vanessa Kuentz-Simonet, Amaury Labenne,Tina Rambonilaza. Using ClustOfVar coupled with social indicators to capture the vulnerability of rural municipalities in southwest France to global change. A paraître dans *Social Indicators Research.*

## Articles soumis

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. The new hclustgeo method to perform hierarchical ascendant clustering with geographical constraints.

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. Multivariate analysis of mixed data : The PCAmixdata R package.

## Conférences

### 2015

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. Le package ClustGeo : Classification ascendante hiérarchique avec contraintes de proximité géographique.*4èmes Rencontres R, Juin 2015, Grenoble.*

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. ClustGeo : Classification ascendante hiérarchique avec contraintes de proximité géographique. *47èmes Journées de Statistique de la SFdS, Juin 2015, Lille.*
— Chavent, M., Kuentz, V., Labenne, A., Rambonilaza, T., Saracco, J. La méthode ClustOfVar pour mesurer les conditions de vie à l'échelle communale. Le complexe littoral / estuaire de la Gironde. *12èmes Journées de Méthodologie Statistique de l'INSEE, Avril 2015, Paris.*

### 2014

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. Variable selection to construct indicators of quality of life for data structured in groups. *COMPSTAT'14, Geneva, Switzerland.*
— Chavent, M., Kuentz, V., Labenne, A., Saracco, J.Une nouvelle approche statistique pour la construction d'indicateurs composites de qualité de vie à l'échelle communale. *3èmes Rencontres R, Juin 2014, Montpellier.*

### 2013

— Chavent, M., Kuentz, V., Labenne, A., Saracco, J. Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes : MFAmix. *5èmes Rencontres des Jeunes Statisticiens, Aout 2013, Aussois.*
— Chavent, M., Kuentz, V., Labenne, A., Rambonilaza, T., Saracco, J. Une extension de l'Analyse Factorielle Multiple pour des groupes de variables mixtes. *2èmes Rencontres R, Juin 2013, Lyon.*
— Chavent, M., Kuentz, V., Labenne, A., Rambonilaza, T., Saracco, J. Une extension de l'Analyse Factorielle Multiple pour des groupes de variables mixtes. *45èmes Journées de Statistique. Mai 2013, Toulouse.*
— Chavent, M., Kuentz, V., Labenne, A., Rambonilaza, T., Saracco, J. Une approche par classification de variables à l'aide de la méthode 'ClustOfVar' pour analyser la qualité de vie à l'échelle communale. *Statlearn'2013. 8-9 avril 2013, Bordeaux.*

## Développement Logiciel

— ClustGeo : Clustering of Observations with Geographical Constraints
Marie Chavent, Vanessa Kuentz, Amaury Labenne and Jerome Saracco (2015). R package version 1.0.

— ClustOfVar : Classification de variables quantitatives et/ou qualitatives.
Marie Chavent, Vanessa Kuentz, Amaury Labenne, Benoit Liquet and Jerome Saracco (2015). ClustOfVar : Clustering of variables. R package version 1.1.

— PCAmixdata : Analyse factorielle de données mixtes (quantitatives / qualitatives).
Marie Chavent, Vanessa Kuentz, Amaury Labenne, Benoit Liquet and Jerome Saracco. (2014). PCAmixdata : Multivariate Analysis of Mixed Data. R package version 2.2.

# Table des figures

# Liste des tableaux

# A

# Ecriture de l'ACM comme une ACP simple

Cette annexe est dédiée à la preuve de l'équation 2.2.13. Nous allons démontrer ici comment réaliser l'ACM à partir d'une seule ACP avec métriques. Initialement, l'ACM est définie comme l'analyse des correspondances appliquée à la matrice des indicatrices $\mathbf{G}$. On note $\mathbf{F}$ la matrice des coordonnées factorielles des observations et $\mathbf{A}^*$ la matrice des coordonnées factorielles des modalités. Les coordonnées factorielles des observations et des modalités sont obtenues en appliquant deux ACP, une appliquée à la matrice des profils-lignes et une autre appliquée à la matrice des profils colonnes. Ces profils lignes et colonnes sont obtenus à l'aide de la matrice des fréquences $\frac{\mathbf{G}}{np}$. Les marges de cette matrice sont utilisées pour pondérer les lignes et les colonnes dans ces deux ACP. On introduit les notations suivantes :

- $\mathbf{r} \in \mathbb{R}^n$ est le vecteur des poids des observations. Le poids des observations est constant et vaut : $\frac{1}{n}$,
- $\mathbf{c} \in \mathbb{R}^m$ est le vecteur des poids des modalités. Le poids d'une modalité $s$ est égal à $\frac{n_s}{np}$,
- $\mathbf{D}_r = \text{diag}(\mathbf{r}) = \frac{1}{n}\mathbb{I}_n$ est la matrice diagonale des poids des observations,
- $\mathbf{D}_c = \text{diag}(\mathbf{c}) = \text{diag}(\frac{n_s}{np}, s = 1, \ldots, m)$ est la matrice diagonale ides poids des modalités,
- $\mathbf{L} = \mathbf{D}_r^{-1}\frac{\mathbf{G}}{np} = \frac{\mathbf{G}}{p}$ est la matrice des profils lignes,
- $\mathbf{C} = \frac{\mathbf{G}}{np}\mathbf{D}_c^{-1}$est la matrice des profils colonnes.

**ACP des lignes de L.** La GSVD de $\mathbf{L}$ avec les métriques $\mathbf{D}_r$ et $\mathbf{D}_c^{-1}$ donne la décomposition suivante :

$$\mathbf{L} = \mathbf{U}_L\mathbf{\Lambda}{\mathbf{V}_L}^t, \tag{A.0.1}$$

où $\mathbf{V}$ est la matrice des vecteurs propres de $\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$ et $\mathbf{U}_L$ est la matrice des vecteurs propres de $\mathbf{L}\mathbf{D}_c^{-1}\mathbf{L}^t\mathbf{D}_r$. Cela donne :

$$\mathbf{F} = \mathbf{L}\mathbf{D}_c^{-1}\mathbf{V}_L = \mathbf{U}_L\mathbf{\Lambda},$$
$$\mathbf{A}_L = \mathbf{L}^t\mathbf{D}_r\mathbf{U}_L = \mathbf{V}_L\mathbf{\Lambda},$$

où $\mathbf{F}$ est la matrice des coordonnées factorielles des lignes de $\mathbf{L}$ et $\mathbf{A}_L$ est la matrice des coordonnées factorielles des colonnes de $\mathbf{L}$. On notera que les colonnes de $\mathbf{L}$ ne sont pas les profils colonnes.

**ACP des colonnes de C.** La GSVD de $\mathbf{C}$ avec les métriques $\mathbf{D}_r^{-1}$ et $\mathbf{D}_c$ donne la décomposition suivante :

$$\mathbf{C} = \mathbf{U}_C\mathbf{\Lambda}{\mathbf{V}_C}^t, \tag{A.0.2}$$

où $\mathbf{U}_C$ est la matrice des vecteurs propres de $\mathbf{C}\mathbf{D}_c\mathbf{C}^t\mathbf{D}_r^{-1}$ et $\mathbf{V}_C$ est la matrice des vecteurs propres de $\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c$. Cela donne :

$$\mathbf{A}^* = \mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{U}_C = \mathbf{V}_C\mathbf{\Lambda},$$
$$\mathbf{F}_C = \mathbf{C}\mathbf{D}_c\mathbf{V}_C = \mathbf{U}_C\mathbf{\Lambda},$$

où $\mathbf{A}^*$ est la matrice des coordonnées factorielles des colonnes de $\mathbf{C}$ et $\mathbf{F}_C$ est la matrice des coordonnées factorielles des lignes de $\mathbf{C}$.

**Une seule ACP de la matrice G.** A ce niveau, les matrices $\mathbf{F}$ et $\mathbf{A}^*$ sont calculées à partir de deux ACP différentes. Nous montrons ci dessous que $\mathbf{V}_C = \mathbf{D}_c^{-1}\mathbf{V}_L$. Il suit que $\mathbf{A}^* = \mathbf{D}_c^{-1}\mathbf{V}_L\mathbf{\Lambda}$ est obtenue directement avec la GSVD de $\mathbf{L}$.

**Proof.** Démontrons tout d'abord que $\mathbf{V}_C = \mathbf{D}_c^{-1}\mathbf{V}_L$. Les matrices $\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$ et $\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c$ ont les mêmes valeurs propres. Soit $\Lambda$ la matrice diagonale de ces valeurs propres. $\mathbf{V}_L$ est la matrice des vecteurs propres de $\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$, on obtient donc :

$$(\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1})\mathbf{V}_L = \Lambda\mathbf{V}_L. \tag{A.0.3}$$

$\mathbf{V}_C$ est la matrice des vecteurs propres de $\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c$, on a donc :

$$(\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c)\mathbf{V}_C = \Lambda\mathbf{V}_C. \tag{A.0.4}$$

$\mathbf{L} = \mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c$. A partir de cette expression de $\mathbf{L}$, nous pouvons écrire $\mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$ en fonction de $\mathbf{C}$ :

$$\begin{aligned} \mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1} &= (\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c)^t\mathbf{D}_r(\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c)\mathbf{D}_c^{-1} \\ &= \mathbf{D}_c\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{D}_r\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c\mathbf{D}_c^{-1} \\ &= \mathbf{D}_c\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}. \end{aligned} \tag{A.0.5}$$

$\mathbf{B} = \mathbf{L}^t\mathbf{D}_r\mathbf{L}\mathbf{D}_c^{-1}$ est $\mathbf{D}_c^{-1}$ symétrique, on a donc :

$$\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{B}^t\mathbf{D}_c^{-1}. \tag{A.0.6}$$

A partir de (A.0.3) on obtient $\mathbf{B}\mathbf{V}_L = \Lambda\mathbf{V}_L$ et donc $\mathbf{D}_c^{-1}\mathbf{B}\mathbf{V}_L = \Lambda\mathbf{D}_c^{-1}\mathbf{V}_L$.
A partir de (A.0.6), on obtient $\mathbf{B}^t\mathbf{D}_c^{-1}\mathbf{V}_L = \Lambda\mathbf{D}_c^{-1}\mathbf{V}_L$.
En utilisant (A.0.5) pour réécrire $\mathbf{B}$, on a :

$$\begin{aligned} (\mathbf{D}_c\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C})^t\mathbf{D}_c^{-1}\mathbf{V}_L &= \Lambda\mathbf{D}_c^{-1}\mathbf{V}_L \\ \text{et } (\mathbf{C}^t\mathbf{D}_r^{-1}\mathbf{C}\mathbf{D}_c)\mathbf{D}_c^{-1}\mathbf{V}_L &= \Lambda\mathbf{D}_c^{-1}\mathbf{V}_L. \end{aligned}$$

Par identification dans (A.0.4), on obtient $\mathbf{V}_C = \mathbf{D}_c^{-1}\mathbf{V}_L$.

■

# Description des variables du jeu de données `gironde` contenu dans le package `PCAmixdata`

La Table B.1 donne la description des variables présentes dans les quatres dataframe de la liste `gironde` présente dans le package `PCAmixdata`.

| R_Names | Description | Group | Data type |
|---|---|---|---|
| farmers | Percentage of farmers | employment | Num |
| tradesmen | Percentage of tradesmen and shopkeepers | employment | Num |
| managers | Percentage of managers and executives | employment | Num |
| workers | Percentage of workers and employees | employment | Num |
| unemployed | Percentage of unemployed workers | employment | Num |
| middleemp | Percentage of middle-range employees | employment | Num |
| retired | Percentage of retired people | employment | Num |
| employrate | Employment rate | employment | Num |
| income | Average income | employment | Num |
| density | Population density | housing | Num |
| primaryres | Percentage of primary residences | housing | Num |
| houses | Percentage of houses | housing | Categ |
| owners | Percentage of home owners living in their primary residence | housing | Num |
| council | Percentage of council housing | housing | Categ |
| butcher | Number of butchers | services | Categ |
| baker | Number of bakers | services | Categ |
| postoffice | Number of post offices | services | Categ |
| dentist | Number of dentists | services | Categ |
| grocery | Number of grocery stores | services | Categ |
| nursery | Number of child care day nurseries | services | Categ |
| doctor | Number of doctors | services | Categ |
| chemist | Number of chemists | services | Categ |
| restaurant | Number of restaurants | services | Categ |
| building | Percentage of buildings | environment | Num |
| water | Percentage of water | environment | Num |
| vegetation | Percentage of vegetation | environment | Num |
| agricul | Percentage of agricultural land | environment | Num |

TABLE B.1 – Description des variables utilisées dans le jeu de données `gironde`

# C

# Description des variables décrivant la zone Garonne-Gironde

Les Tables C.1 et C.2 présentées ci-après détaillent les variables utilisées en 1999 et en 2009 pour la construction des indicateurs composites réalisée grâce à la méthode hclustvar à la Section 3.3. On rappelle que ces variables ont été mesurées sur la zone Garonne-Gironde présentée à la Section 3.3.1. Sauf indication contraire les moyennes sont données en pourcentages.

| Nom de la variable | Description | Moyenne |
|---|---|---|
| ActOqp_99 | Part des actifs occupés dans les actifs totaux | 88 |
| AgrExpl_99 | Part des agriculteurs, exploitants dans l'ensemble CSP | 7 |
| AGRI | Pourcentage de la commune couvert par des terres agricoles | 71 |
| ArtComCE_99 | Part des artisans, commerçants, chefs d'entreprise dans l'ensemble CSP | 4 |
| Banque$^{\diamond}$ | Présence de banques | 12 |
| BATI | Pourcentage de la commune couvert par des batiments | 4 |
| BouChar$^{\diamond}$ | Présence de boucheries, charcuteries | 22 |
| BoulPat$^{\diamond}$ | Présence de boulangeries, patisseries | 33 |
| CadInt_99 | Part des cadres et professions intellectuelles supérieures dans l'ensemble CSP | 4 |
| CafeBoissons$^{\diamond}$ | Présence de bars et débits de boissons | 49 |
| Carburant$^{\diamond}$ | Présence de pompes à essences | 23 |
| College$^{\diamond}$ | Présence de collèges | 8 |
| CrecheColl$^{\diamond}$ | Présence de crèches collectives | 5 |
| CrecheFam$^{\diamond}$ | Présence de crèches familiales | 9 |
| Densite_99 | Densité de population (Habitants/Km2) | 80 |
| DistanceFleuve | Distance entre la mairie et la Garonnne (en mètres) | 23883 |
| EAU | Pourcentage de la commune couvert par des surfaces en eau | 1 |
| EcoleMat$^{\diamond}$ | Présence d'écoles maternelles | 47 |
| EcolePrim$^{\diamond}$ | Présence d'écoles primaires | 38 |
| Emploi1524_99 | Taux d'emploi des 15 à 24 ans dans la population de même tranche d'âge | 23 |
| Emploi2554_99 | Taux d'emploi des 25 à 54 ans dans la population de même tranche d'âge | 78 |
| Emploi5564_99 | Taux d'emploi des 55 à 64 ans dans la population de même tranche d'âge | 34 |

Suite page suivante

| Nom de la variable | Description | Moyenne |
|---|---|---|
| EmploICommune_99 | Part des actifs occupés dans une commune de même département que la commune de résidence | 32 |
| EmploiDpt_99 | Part des actifs occupés dans une commune de même département que la commune de résidence | 57 |
| EmploiUU_99 | Part des actifs occupés dans une commune de même unité urbaine que la commune de résidence | 5 |
| EmploiZE_99 | Part des actifs occupés dans une commune de même zone emploi que la commune de résidence | 49 |
| GardPeriscol$^{\diamond}$ | Présence de gardes d'enfants périscolaire | 46 |
| HalteGard$^{\diamond}$ | Présence de halte garderies | 10 |
| Logmt1548_99 | Part des logements construits entre 1915 et 1948 | 8 |
| Logmt4974_99 | Part des logements construits entre 1949 et 1974 | 14 |
| Logmt7589_99 | Part des logements construits entre 1975 et 1989 | 21 |
| LogmtAp90_99 | Part des logements construits entre 1990 et 1999 | 11 |
| MedGen$^{\diamond}$ | Présence de médecins généralites | 25 |
| Menage_famprinc_cpleavecenfants_99 | Part des ménages dont la famille principale est composée d'un couple avec enfants | 36 |
| Menage_famprinc_cplesansenfants_99 | Part des ménages dont la famille principale est composée d'un couple sans enfant | 32 |
| Menage_famprinc_monoparentale_99 | Part des ménages dont la famille principale est une famille monoparentale | 7 |
| Menage_Fseule_99 | Part des ménages composés d'une femme seule | 12 |
| Menage_Hseul_99 | Part des ménages composés d'un homme seul | 11 |
| NbMoyPieceLog_99 | Nombre moyen de pièces par logements | 4 |
| Ndip_ns_99 | Part de la population non scolarisée de 15 ans et plus titulaire d'aucun diplôme | 21 |
| OuvEmp_99 | Part des ouvriers et employés dans l'ensemble CSP | 28 |
| Pharmacie$^{\diamond}$ | Présence de pharmacies | 20 |
| Poste$^{\diamond}$ | Présence de postes | 30 |
| ProfInter_99 | Part des professions intermédiaires dans l'ensemble CSP | 9 |
| Restaurant$^{\diamond}$ | Présence de restaurants | 41 |
| Ret_99 | Part des retraités dans l'ensemble CSP | 28 |
| RNI_99 | Revenu moyen des foyers fiscaux (en euros) | 8489895 |
| RPHLMLoc_99_Plus_5% $^{\diamond}$ | Présence de plus de 5% de logements HLM | 11 |
| RPOccProp_99 | Part des résidences principales occupées par le propriétaire dans les résidences principales totales | 75 |
| RPTypeMai_99_Plus_90% $^{\diamond}$ | Présence de plus de 90% de logements de type maison | 64 |
| SAP_99 | Part des personnes sans activité professionnelle dans l'ensemble CSP | 20 |
| Supermarche$^{\diamond}$ | Présence de supermarché | 9 |
| Tabac$^{\diamond}$ | Présence de tabacs | 42 |
| VEGE | Pourcentage de la commune couvert par de la végétation | 25 |
| Veterinaire$^{\diamond}$ | Présence de vétérinaires | 8 |

TABLE C.1 – Dictionnaire des variables de l'année 1999 mesurées sur la zone Garonne-Gironde. Le symbole $^{\diamond}$ indique que la variable est qualitative.

| Nom de la variable | Description | Moyenne |
|---|---|---|
| ActOqp_09 | Part des actifs occupés dans les actifs totaux | 91 |
| AgrExpl_09 | Part des agriculteurs, exploitants dans l'ensemble CSP | 5 |
| AGRI | Pourcentage de la commune couvert par des terres agricoles | 70 |
| ArtComCE_09 | Part des artisans, commerçants, chefs d'entreprise dans l'ensemble CSP | 5 |
| BanqueCE$^{\diamond}$ | Présence de banques | 13 |
| BATI | Pourcentage de la commune couvert par des bâtiments | 4 |
| BouChar$^{\diamond}$ | Présence de boucheries, charcuteries | 20 |
| Boul$^{\diamond}$ | Présence de boulangeries, patisseries | 32 |
| CadInt_09 | Part des cadres et professions intellectuelles supérieures dans l'ensemble CSP | 5 |
| ChiDentiste$^{\diamond}$ | Présence de chirurgiens dentistes | 17 |
| College$^{\diamond}$ | Présence de collèges | 8 |
| Densite_09 | Densité de population (Habitants/Km2) | 90 |
| DistanceFleuve | Distance entre la mairie et la Garonnne (en mètres) | 23875 |
| EAU | Pourcentage de la commune couvert par des surfaces en eau | 1 |
| EcolElem$^{\diamond}$ | Présence d'écoles primaires | 32 |
| EcoleMat$^{\diamond}$ | Présence d'écoles maternelles | 15 |
| Emploi1524_09 | Taux d'emploi des 15 à 24 ans dans la population de même tranche d'âge | 34 |
| Emploi2554_09 | Taux d'emploi des 25 à 54 ans dans la population de même tranche d'âge | 83 |
| Emploi5564_09 | Taux d'emploi des 55 à 64 ans dans la population de même tranche d'âge | 38 |
| EmploiComm_09 | Part des actifs occupés dans une commune de même département que la commune de résidence | 26 |
| EmploiDpt_09 | Part des actifs occupés dans une commune de même département que la commune de résidence | 62 |
| Epicerie$^{\diamond}$ | Présence d'épiceries | 22 |
| GardPrescol$^{\diamond}$ | Présence de gardes pré-scolaires | 7 |
| MedOmni$^{\diamond}$ | Présence de médecins omnipraticiens | 24 |
| Menage_famprinc_cpleavecenfants_09 | Part des ménages dont la famille principale est composée d'un couple avec enfants | 31 |
| Menage_famprinc_cplesansenfants_09 | Part des ménages dont la famille principale est composée d'un couple sans enfant | 34 |
| Menage_famprinc_monoparentale_09 | Part des ménages dont la famille principale est une famille monoparentale | 7 |
| Menage_Fseule_09 | Part des ménages composés d'une femme seule | 13 |
| Menage_Hseul_09 | Part des ménages composés d'un homme seul | 12 |
| Ndip_ns_09 | Part de la population non scolarisée de 15 ans et plus titulaire d'aucun diplôme | 17 |
| OuvEmp_09 | Part des ouvriers et employés dans l'ensemble CSP | 29 |
| Pharmacie$^{\diamond}$ | Présence de pharmacies | 20 |
| Poste$^{\diamond}$ | Présence de postes | 24 |
| ProfInter_09 | Part des professions intermédiaires dans l'ensemble CSP | 12 |
| Restaurant$^{\diamond}$ | Présence de restaurants | 39 |
| Ret_09 | Part des retraités dans l'ensemble CSP | 32 |
| RNIMoy_09 | Revenu moyen des foyers fiscaux (en euros) | 20277 |
| RPHLMLoc_09_Plus_5% $^{\diamond}$ | Présence de plus de 5% de logements HLM | 12 |
| RPLogTot_09 | Part des résidences principales parmi l'ensemble des logements | 79 |
| RPOccProp_09 | Part des résidences principales occupées par le propriétaire dans les résidences principales totales | 78 |
| RPTypeMai_09_Plus_90% $^{\diamond}$ | Présence de plus de 90% de logements de type maison | 81 |
| SAP_09 | Part des personnes sans activité professionnelle dans l'ensemble CSP | 12 |

Suite page suivante

| Nom de la variable | Description | Moyenne |
|---|---|---|
| Superette$^{\diamond}$ | Présence de superettes | 6 |
| Supermarche$^{\diamond}$ | Présence de supermarché | 10 |
| VEGE | Pourcentage de la commune couvert par de la végétation | 25 |
| Veterinaire$^{\diamond}$ | Présence de vétérinaires | 10 |

TABLE C.2 – Dictionnaire des variables de l'année 2009 mesurées sur la zone Garonne-Gironde. Le symbole $^{\diamond}$ indique que la variable est qualitative.

# D

# Description des variables utilisées sur la zone des SAGEs

La Table D.1 donne la description des variables utilisées à la Section 4.4 pour construire des indicateurs sur la zone des SAGESs à l'aide de la méthode MFAmix.

| Groupe dans MFAmix | Nom de la variable | Description de la variable |
|---|---|---|
| Logement | RPLogTot | Part des résidences principales dans les logements totaux en 2009 (%) |
| | RPTypeMai | Part des résidences principales de type maison dans les résidences principales en 2009 (%) |
| | RPOccProp | Part des résidences principales occupées par le propriétaire dans les résidences principales totales en 2009 (%) |
| | RPHLMLoc$^{\diamond}$ | Part des résidences principales HLM en location dans les résidences principales totales en 2009 (%) |
| Emploi | ActOqp | Part des actifs occupés dans les actifs totaux en 2009 (%) |
| | Emploi1524 | Taux d'emploi des 15 à 24 ans dans la population de même tranche d'âge en 2009 (%) |
| | Emploi2554 | Taux d'emploi des 25 à 54 ans dans la population de même tranche d'âge en 2009 (%) |
| | Emploi5564 | Taux d'emploi des 55 à 64 ans dans la population de même tranche d'âge en 2009 (%) |
| | EmploiDpt | Part des actifs occupés dans une commune de même département que la commune de résidence en 2009 (%) |
| | EmploiComm | Part des actifs occupés travaillant dans leur commune de résidence en 2009 (%) |
| Niveaux de vie | AgrExpl | Part des agriculteurs, exploitants dans l'ensemble CSP en 2009 (%) |
| | ArtComCE | Part des artisans, commerçants, chefs d'entreprise dans l'ensemble CSP en 2009 (%) |
| | CadInt | Part des cadres et professions intellectuelles supérieures dans l'ensemble CSP en 2009 (%) |
| | OuvEmp | Part des ouvriers et employés dans l'ensemble CSP en 2009 (%) |
| | ProfInter | Part des professions intermédiaires dans l'ensemble CSP en 2009 (%) |
| | RNIMoy | Revenu moyen des foyers fiscaux en 2009 (en euros) |
| | Ndip_ns | Part de la population non scolarisée de 15 ans et plus titulaire d'aucun diplôme en 2009 (%) |

Suite page suivante

| Groupe dans MFAmix | Nom de la variable | Description de la variable |
|---|---|---|
| Modes de vie | SAP | Part des personnes sans activité professionnelle dans l'ensemble CSP en 2009 (%) |
| | Ret | Part des retraités dans l'ensemble CSP en 2009 (%) |
| | Densite | Densité de population en 2009 (Habitants/Km2) |
| | Menage_famprinc_cpleavecenfants | Part des ménages dont la famille principale est composée d'un couple avec enfants en 2009 |
| | Menage_famprinc_cplesansenfants | Part des ménages dont la famille principale est composée d'un couple sans enfant en 2009 (%) |
| | Menage_famprinc_monoparentale | Part des ménages dont la famille principale est une famille monoparentale en 2009 (%) |
| | Menage_Fseule | Part des ménages composés d'une femme seule en 2009 (%) |
| | Menage_Hseul | Part des ménages composés d'un homme seul en 2009 (%) |
| Environnement | Bati | Pourcentage de la commune couvert par des batiments (%) |
| | EAU | Pourcentage de la commune couvert par des surfaces en eau (%) |
| | Vegetation | Pourcentage de la commune couvert par de la végétation (%) |
| | Agri | Pourcentage de la commune couvert par des terres agricoles (%) |
| Services | BanqueCE$^{\diamond}$ | Banque, Caisse d'Epargne |
| | BouChar$^{\diamond}$ | Boucherie, charcuterie |
| | Boul$^{\diamond}$ | Boulangerie |
| | Poste$^{\diamond}$ | Bureau de poste |
| | ChiDentiste$^{\diamond}$ | Chirurgien dentiste |
| | College$^{\diamond}$ | Collège |
| | EcolElem$^{\diamond}$ | Ecole élémentaire |
| | EcoleMat$^{\diamond}$ | Ecole maternelle |
| | Epicerie$^{\diamond}$ | Epicerie |
| | GardPrescol$^{\diamond}$ | Garde d'enfant d'âge préscolaire |
| | MedOmni$^{\diamond}$ | Médecin omnipraticien |
| | Pharmacie$^{\diamond}$ | Pharmacie |
| | Restaurant$^{\diamond}$ | Restaurant |
| | Superette$^{\diamond}$ | Superette |
| | Supermarche$^{\diamond}$ | Supermarché |
| | Veterinaire$^{\diamond}$ | Vétérinaire |

TABLE D.1 – Dictionnaire des variables de l'année 2009 mesurées sur la zone des SAGEs. Le symbole $^{\diamond}$ indique que la variable est qualitative

# Bibliographie

Ambroise, C., M. Dang, and G. Govaert (1997). Clustering of Spatial Data by the EM Algorithm. In A. Soares, J. Gómez-Hernandez, and R. Froidevaux (Eds.), *geoENV I — Geostatistics for Environmental Applications*, Quantitative Geology and Geostatistics, pp. 493–504. Springer Netherlands.

Bécue-Bertaut, M. and J. Pagès (2008, February). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis 52*(6), 3255–3268.

Beaton, D., C. R. Chin Fatt, and H. Abdi (2014, April). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis 72*, 176–189.

Beh, E. J. and R. Lombardo (2012). A Genealogy of Correspondence Analysis. *Australian & New Zealand Journal of Statistics 54*(2), 137–168.

Besse, P. (1992, April). PCA stability and choice of dimensionality. *Statistics & Probability Letters 13*(5), 405–410.

Bovar, O. and F. Nirascou (2010). Des indicateurs du développement durable pour les territoires. *La Revue*, 43–54.

Chavent, M., V. Kuentz, A. Labenne, B. Liquet, and J. Saracco (2014). PCAmixdata : Multivariate Analysis of Mixed Data. R package version 2.2.

Chavent, M., V. Kuentz, A. Labenne, B. Liquet, and J. Saracco (2015). Clustofvar : Clustering of variables. R package version 1.1.

Chavent, M., V. Kuentz Simonet, B. Liquet, and J. Saracco (2012). ClustOfVar : An R Package for the Clustering of Variables. *Journal of Statistical Software 50*(13), 1–16.

Chavent, M., V. Kuentz-Simonet, and J. Saracco (2012, March). Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classification 6*(2), 131–146.

Chavent, M., Y. Lechevallier, F. Vernier, and K. Petit (2008). Monothetic Divisive Clustering with Geographical Constraints. In P. Brito (Ed.), *COMPSTAT 2008*, pp. 67–76. Physica-Verlag HD.

Costanza, R., B. Fisher, S. Ali, C. Beer, L. Bond, R. Boumans, N. L. Danigelis, J. Dickinson, C. Elliott, J. Farley, D. E. Gayer, L. M. Glenn, T. Hudspeth, D. Mahoney, L. McCahill, B. McIntosh, B. Reed, S. A. T. Rizvi, D. M. Rizzo, T. Simpatico, and R. Snapp (2007, March). Quality of life : An approach integrating opportunities, human needs, and subjective well-being. *Ecological Economics 61*(2–3), 267–276.

Coudret, R., B. Liquet, and J. Saracco (2014, December). Comparison of sliced inverse regression approaches for underdetermined cases. *Journal de la Société Française de Statistique 155*(2), 72–96.

Dhillon, I. S., E. M. Marcotte, and U. Roshan (2003, September). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics 19*(13), 1612–1619.

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'analyse des données 4*(2), 137–146.

Escofier, B. and J. Pagès (1983). Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire. *Revue de Statistique Appliquée 31*(2), 43–59.

Giraud, T. (2013). rCarto : This package builds maps with a full cartographic layout. R package version 0.8.

Gordon, A. D. (1996, January). A survey of constrained classification. *Computational Statistics & Data Analysis 21*, 17–29.

Guo, D. (2009, December). Greedy Optimization for Contiguity-Constrained Hierarchical Clustering. In *IEEE International Conference on Data Mining Workshops, 2009. ICDMW '09*, pp. 591–596.

Haines-Young, R. and M. Potschin (2010). Proposal for a common international classification of ecosystem goods and services (cices) for integrated environmental and economic accounting. *Report to the European Environment Agency*.

Haq, R. and Zia (2013). Multidimensional wellbeing : An index of quality of life in a developing economy. *Social Indicators Research 114*, 997–1012.

Hill, M. O. and A. J. E. Smith (1976, May). Principal Component Analysis of Taxonomic Data with Multi-State Discrete Characters. *Taxon 25*(2/3), 249–255.

Husson, F. and J. Josse (2013, December). Handling missing values in multiple factor analysis. *Food Quality and Preference 30*(2), 77–85.

Husson, F., J. Josse, S. Le, and J. Mazet (2015). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining*. R package version 1.29.

Josse, J. and F. Husson (2012, June). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis 56*(6), 1869–1879.

Josse, J., J. Pagès, and F. Husson (2008, September). Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis 53*(1), 82–91.

Kaiser, H. F. (1958, September). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187–200.

Kiers, H. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika 56*(2), 197–212.

Krishnan, V. (2014). Development of a multidimensional living conditions index. *Social Indicators Research*, 1–27.

Lance, G. N. and W. T. Williams (1967, February). A General Theory of Classificatory Sorting Strategies 1. Hierarchical Systems. *The Computer Journal 9*, 373–380.

Legendre, P. and L. Legendre (2012). Chapter 12 - Ecological data series. In P. L. a. L. Legendre (Ed.), *Developments in Environmental Modelling*, Volume 24 of *Numerical Ecology*, pp. 711–783. Elsevier.

Murtagh, F. (1985, January). A Survey of Algorithms for Contiguity-constrained Clustering and Related Problems. *The Computer Journal 28*, 82–88.

Nardo (2008, August). *Handbook on Constructing Composite Indicators : Methodology and User Guide*. OECD Publishing.

Noll, H.-H. (2002, June). Towards a European System of Social Indicators : Theoretical Framework and System Architecture. *Social Indicators Research 58*(1-3), 47–87.

Oliver, M. and R. Webster (1988). A geostatistical basis for spatial weighting in multivariate classification. *21*, 15–35.

Pagès, J. (1996). Eléments de comparaison entre l'Analyse Factorielle Multiple et la méthode STATIS. *Revue de statistique appliquée 44*(4), 81–95.

Pagès, J. (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée 50*(4), 5–37.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée 52*(4), 93–111.

Reynard, R. and P. Vialette (2014). Une approche de la qualité de vie dans les territoires. *Insee Premiere 1519*.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions TECHNIP.

Vigneau, E. and M. Chen (2015). Clustvarlv : Clustering of variables around latent variables. R package version 1.4.1.

Vigneau, E. and E. M. Qannari (2003, January). Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation 32*(4), 1131–1150.

Vigneau, E., E. M. Qannari, K. Sahmer, and D. Ladiray (2006). Classification de variables autour de composantes latentes. *Revue de statistique appliquée 54*(1), 27–45.

Webster, R. (1977). *Quantitative and numerical methods in soil classification and survey*. Oxford ; New York : Clarendon Press.

Webster, R. and P. A. Burrough (1972, June). Computer-Based Soil Mapping of Small Areas from Sample Data. *Journal of Soil Science 23*, 222–234.

Wood, N., C. Burton, and S. Cutter (2010). Community variations in social vulnerability to cascadia-related tsunamis in the u.s. pacific northwest. *Natural Hazards 52*, 369–389.